Title Goes Here

Yichun Chen1,✉, and Ben Marwick2

23 March, 2020

Text of abstract

1 University of One Place  
2 University of Another Place

✉ Correspondence: [Yichun Chen <[yichun33@uw.edu](mailto:yichun33@uw.edu)>](mailto:yichun33@uw.edu)

# Introduction

Here is a citation (Marwick, 2017)

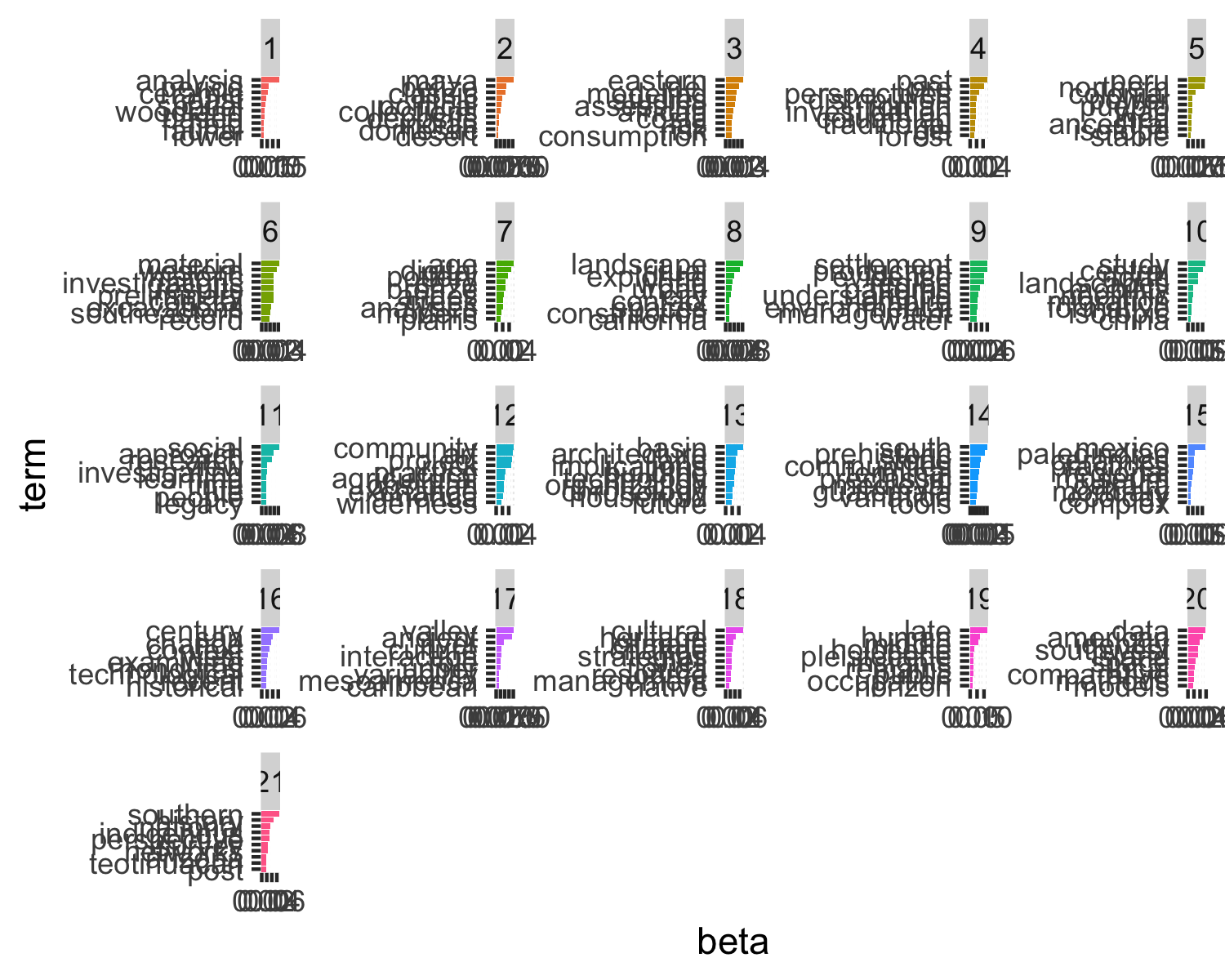
Gender under-representation and inequality has long existed, from under-representation in children’s picture books to academic publications (Hamilton et al. 2006; Tushingham et al. 2017). When talking about inequality in academic fields, we usually refer to the disproportionality between men and women in the field, including archaeology. Women are often underrepresented due to the lack of support women receive in academic fields (Xu 2008). When focusing on the field of archaeology, we can see that gender imbalance still occurs in archaeology and that seems to impact the kind of publications that are done in the field (Bardolph et al. 2016). As we can see in the authorship of the main authors from the Journal of Field Archaeology, 78% of them are men (Heath-Stout 2020). When looking at gender ratios among the membership in Society for California Archaeology from 1967 to 2016, Tushingham et al. (2017) report a trend of women increasingly maintaining their society membership and the gender gap is almost gone. In this study we investigate gender ratios of presenters at major archaeological conferences. We extend the work of Tushingham et al. (2017) with a study of gender ratios of presenters at meetings of the Society of American Archaeology (SAA, 2016-2019), the European Association of Archaeologists (EAA, 2018) and the Computer Applications and Quantitative Methods in Archaeology (CAA, 2017) We explore how gender ratios vary over time and between these conferences. We examine the influence of the gender of the presenter on the topics that they present on.

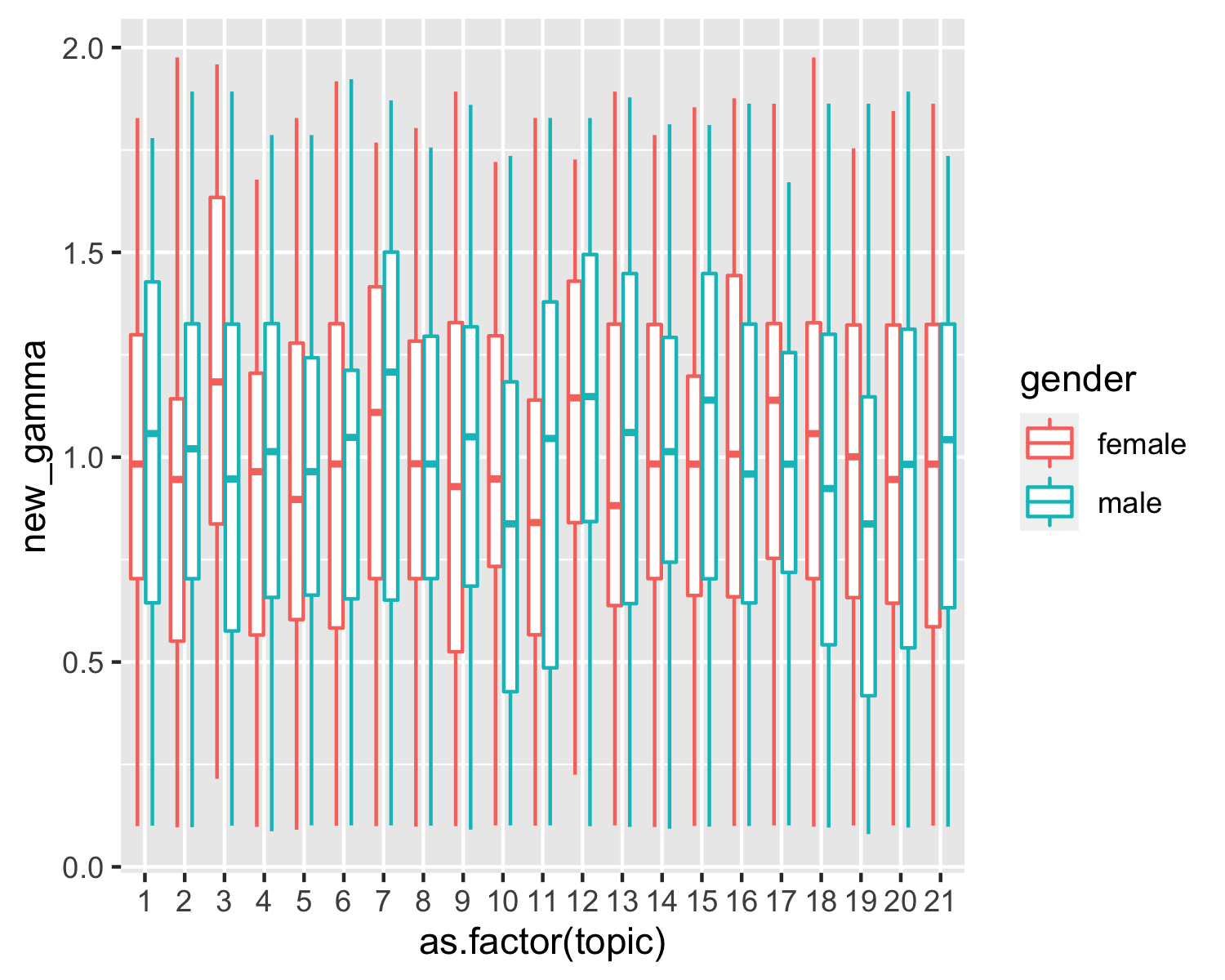
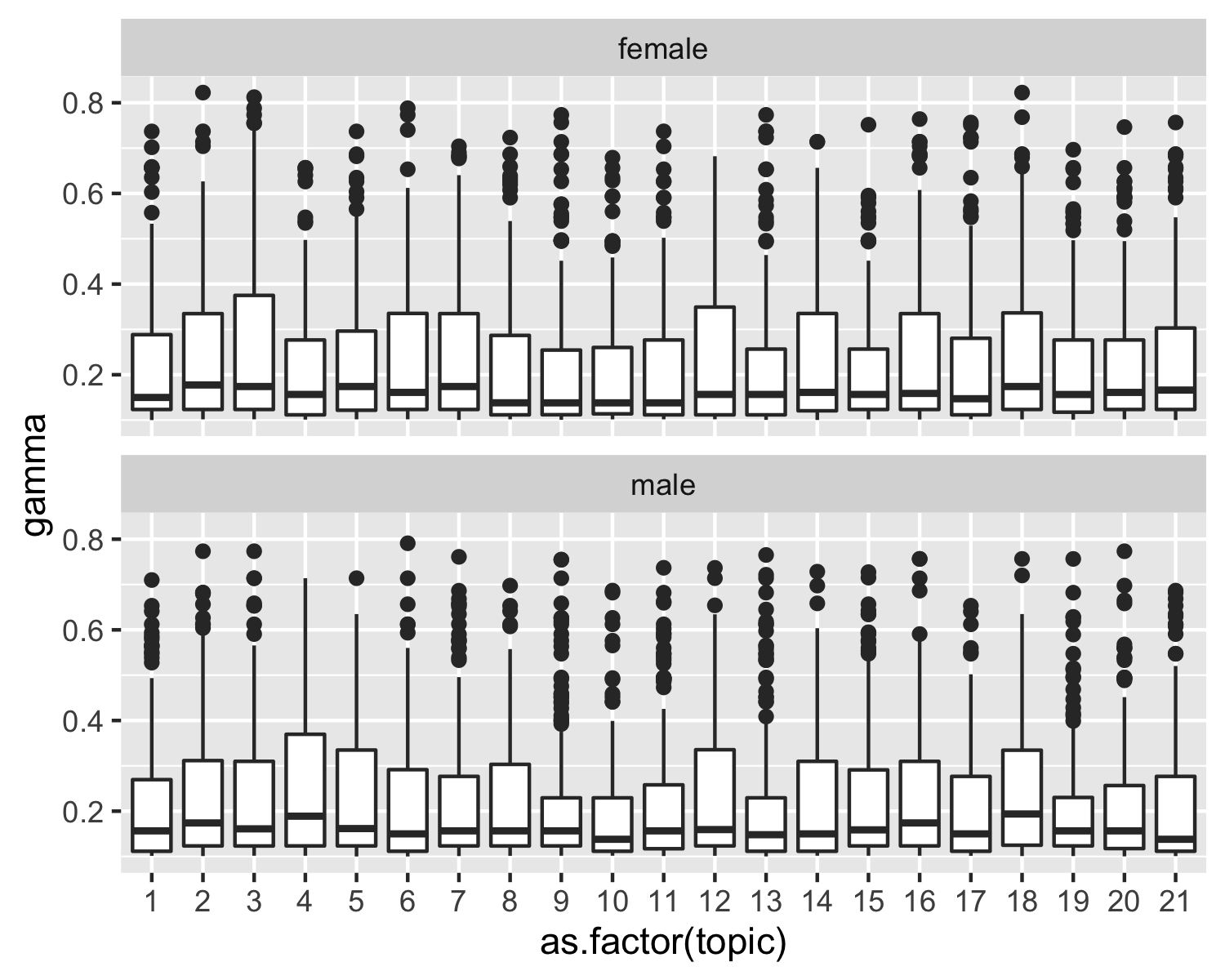
# Background

# Methods

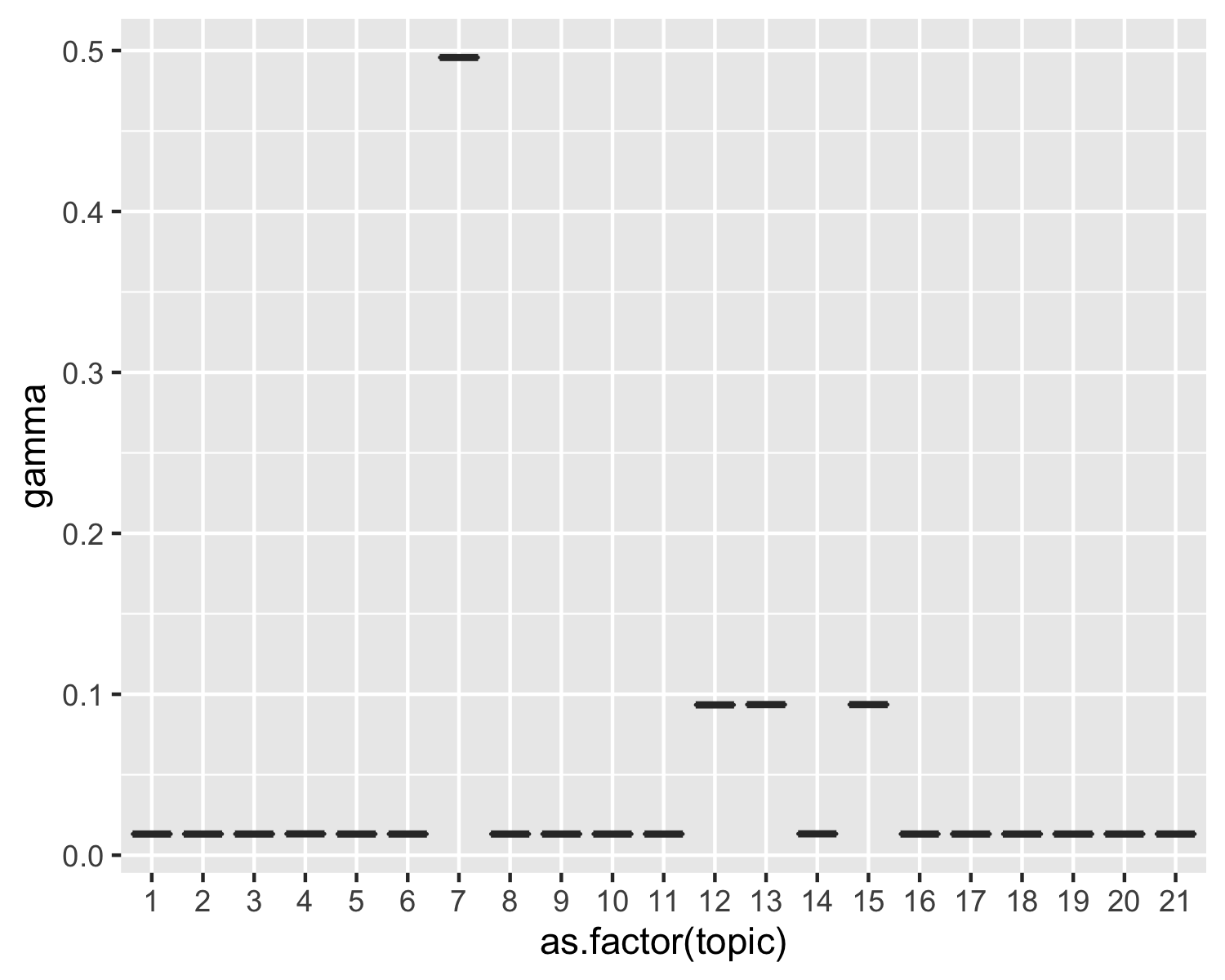
We collected lists of the people that presented research at recent SAA (2016 with 905 presenters, 2017 with 992 presenters, 2018 with 3512 presenters, 2019 with 1014 presenters) EAA (2018 with 2928 presenters) and CAA (2018 with 358 presenters) conferences. We obtained publicly available programs of these conferences and also tidied the data into a rectangular form with variables including the primary author’s first name, last name and the title of their presentation. We used R and the gender package (Mullen 2018; Mihaljević 2019; Belvis et al. 2015) to predict the gender of all first-named presenters in our sample. We used the gender package to identify a first name as male or female, based on a probability scale computed from government records of name-gender data, specifically US Social Security Administration (SSA) baby name data [citation]. This method has the advantage of speed and ease of automation, but also some substantial limitations. We are only able to infer gender into binary male/female genders and assign probabilities for the first names into these categories. This has the unfortunate result of excluding other genders from the results, and we encourage conference organizers to collect gender data directly from participants to improve the representation of minority genders in future studies. The package often fails to classify non-English names, as the SSA data the package uses are mostly in English names. This means that people with non-English names are underrepresented in our results. To remove names where gender is ambiguous, which we define as a similar proportion of males and females having that name in the SSA data, we excluded names where the difference in proportions was less than 0.47, a figure we obtained by exploratory data analysis. It is important to note that the gender designations here are not self-identified by the presenters, but are computed probabilistically. Better quality data will come if presenters self-identify their gender, but we have no alternative because currently these data are not available. Our hope is that this work, despite its many limitations, will stimulate the collection of more reliable, justifiable, and useful data by conference organisers.

We are also interested in whether there is a variation in topic choices between the two biological genders in archaeological conferences. To do this we have used the process of topic modeling, which is used to find natural groups of words when we do not have a specific criteria or simply unsure what we are looking for. There are many ways to accomplish topic modeling, we have used the method of Latent Dirichlet allocation (LDA). This method is beneficial in our case because LDA does not treat one topic, which consists of many words, as separate from another topic. Before we can start using LDA, we have to first decide on the number of topics in our collection of texts. To do this we have used the R package ldatuning (add citation), which allows us to use multiple methods to compute an optimal number of groups for LDA processing. We then create a topic model for all our texts, this will assign all words in every document into topics, and assign each word a probability based on its per-topic-per-word probability, which we call beta. After this we keep only the top 10 words within each topic as keywords to represent each topic. Below is an example of a plot for the 2018 SAA that shows the predetermined number of groups each with top ten keywords in each group. From this we then have top terms from the topics of all conferences. We are only interested in keywords that are meaningful in a topic, therefore we have excluded words that are common in the field of archaeology or simply meaningless, for example, the, a, el, al, archaeology…etc. After figuring out the top terms, we can combine them with our earlier exploration of gender to see if one gender seems to favor another, for example if female or male favor topics about ‘rock’.





#> Df Sum Sq Mean Sq F value Pr(>F)   
#> gender 1 0.00 0.002773 0.149 0.7000   
#> topic 20 0.60 0.029920 1.602 0.0438 \*  
#> gender:topic 20 0.55 0.027641 1.480 0.0777 .  
#> Residuals 2460 45.93 0.018671   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



#>   
#> Welch Two Sample t-test  
#>   
#> data: gamma by gender  
#> t = -2.2562, df = 297.16, p-value = 0.02478  
#> alternative hypothesis: true difference in means is not equal to 0  
#> 95 percent confidence interval:  
#> -0.072788004 -0.004966633  
#> sample estimates:  
#> mean in group female mean in group male   
#> 0.2208258 0.2597032

# Results

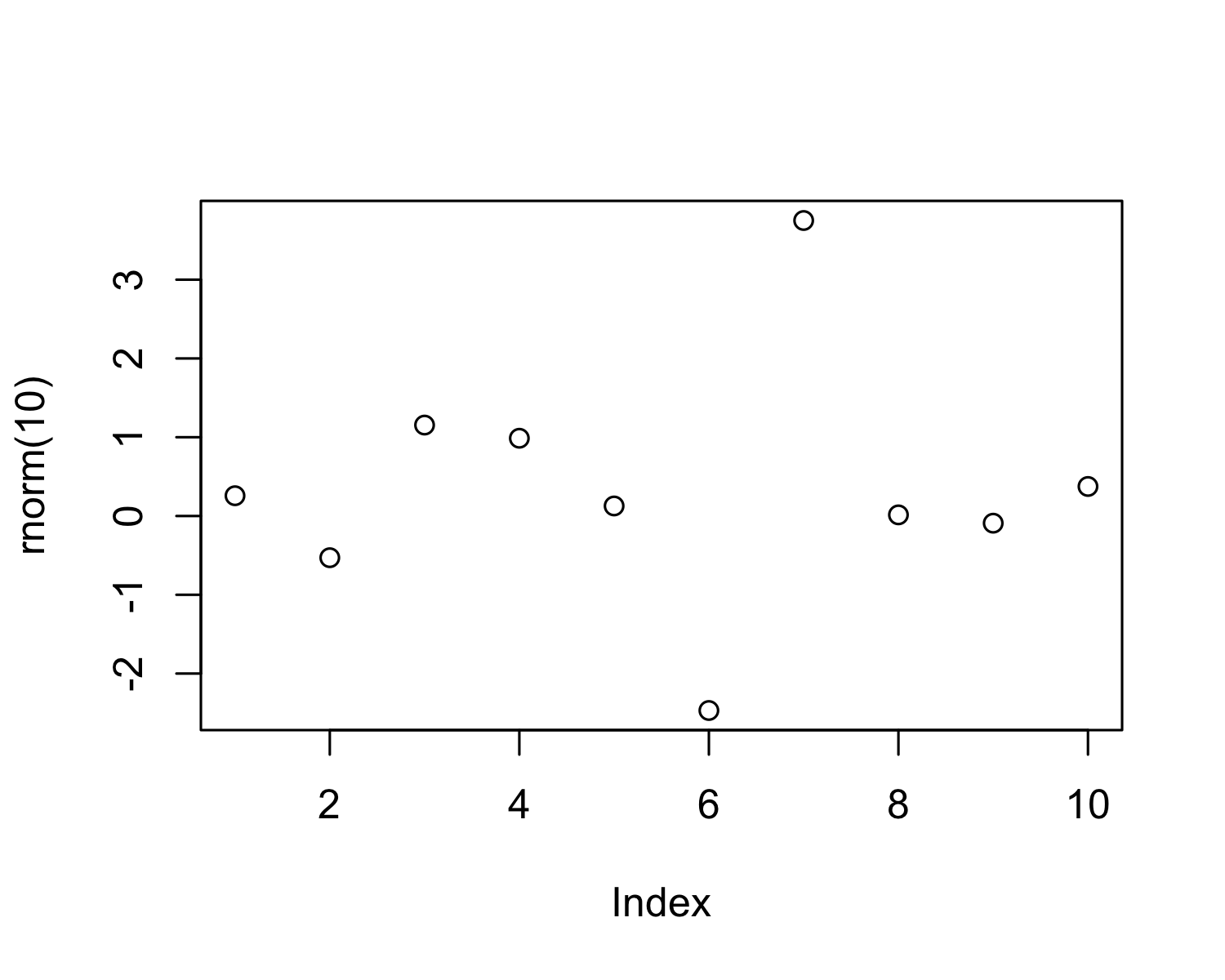


Figure 1: A plot of random numbers

Figure 1 shows how we can have a caption and cross-reference for a plot

Here is an example of inline code 3.14 in the middle of a sentence.

# Discussion

# Conclusion

# Acknowledgements

# References

Marwick, B., 2017. Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. Journal of Archaeological Method and Theory 24, 424–450. <https://doi.org/10.1007/s10816-015-9272-9>

### Colophon

This report was generated on 2020-03-23 17:22:06 using the following computational environment and dependencies:

#> ─ Session info ───────────────────────────────────────────────────────────────  
#> setting value   
#> version R version 3.6.0 (2019-04-26)  
#> os macOS 10.15.2   
#> system x86\_64, darwin15.6.0   
#> ui X11   
#> language (EN)   
#> collate en\_US.UTF-8   
#> ctype en\_US.UTF-8   
#> tz America/Los\_Angeles   
#> date 2020-03-23   
#>   
#> ─ Packages ───────────────────────────────────────────────────────────────────  
#> package \* version date lib source   
#> acepack 1.4.1 2016-10-29 [1] CRAN (R 3.6.0)   
#> assertthat 0.2.1 2019-03-21 [1] CRAN (R 3.6.0)   
#> backports 1.1.5 2019-10-02 [1] CRAN (R 3.6.0)   
#> base64enc 0.1-3 2015-07-28 [1] CRAN (R 3.6.0)   
#> bookdown 0.18 2020-03-05 [1] CRAN (R 3.6.0)   
#> broom 0.5.5 2020-02-29 [1] CRAN (R 3.6.0)   
#> callr 3.4.2 2020-02-12 [1] CRAN (R 3.6.0)   
#> cellranger 1.1.0.9000 2019-05-28 [1] Github (rsheets/cellranger@7ecde54)  
#> checkmate 2.0.0 2020-02-06 [1] CRAN (R 3.6.0)   
#> cli 2.0.2 2020-02-28 [1] CRAN (R 3.6.0)   
#> cluster 2.1.0 2019-06-19 [1] CRAN (R 3.6.0)   
#> colorspace 1.4-1 2019-03-18 [1] CRAN (R 3.6.0)   
#> crayon 1.3.4 2020-01-31 [1] Github (gaborcsardi/crayon@f4bc7b8)  
#> data.table 1.12.8 2019-12-09 [1] CRAN (R 3.6.0)   
#> DBI 1.1.0 2019-12-15 [1] CRAN (R 3.6.0)   
#> dbplyr 1.4.2 2019-06-17 [1] CRAN (R 3.6.0)   
#> desc 1.2.0 2018-05-01 [1] CRAN (R 3.6.0)   
#> devtools 2.2.2 2020-02-17 [1] CRAN (R 3.6.0)   
#> digest 0.6.25 2020-02-23 [1] CRAN (R 3.6.0)   
#> dplyr \* 0.8.5 2020-03-07 [1] CRAN (R 3.6.0)   
#> ellipsis 0.3.0 2019-09-20 [1] CRAN (R 3.6.0)   
#> evaluate 0.14 2019-05-28 [1] CRAN (R 3.6.0)   
#> fansi 0.4.1 2020-01-08 [1] CRAN (R 3.6.0)   
#> farver 2.0.3 2020-01-16 [1] CRAN (R 3.6.0)   
#> forcats \* 0.5.0 2020-03-01 [1] CRAN (R 3.6.0)   
#> foreign 0.8-76 2020-03-03 [1] CRAN (R 3.6.0)   
#> Formula \* 1.2-3 2018-05-03 [1] CRAN (R 3.6.0)   
#> fs 1.3.2 2020-03-05 [1] CRAN (R 3.6.0)   
#> gender \* 0.5.3 2019-11-09 [1] CRAN (R 3.6.0)   
#> genderdata \* 0.5.0 2020-03-23 [1] local   
#> generics 0.0.2 2018-11-29 [1] CRAN (R 3.6.0)   
#> ggplot2 \* 3.3.0 2020-03-05 [1] CRAN (R 3.6.0)   
#> glue 1.3.2 2020-03-12 [1] CRAN (R 3.6.0)   
#> gridExtra 2.3 2017-09-09 [1] CRAN (R 3.6.0)   
#> gtable 0.3.0 2019-03-25 [1] CRAN (R 3.6.0)   
#> haven 2.2.0 2019-11-08 [1] CRAN (R 3.6.0)   
#> here \* 0.1 2017-05-28 [1] CRAN (R 3.6.0)   
#> highr 0.8 2019-03-20 [1] CRAN (R 3.6.0)   
#> Hmisc \* 4.3-1 2020-02-07 [1] CRAN (R 3.6.0)   
#> hms 0.5.3 2020-01-08 [1] CRAN (R 3.6.0)   
#> htmlTable 1.13.3 2019-12-04 [1] CRAN (R 3.6.0)   
#> htmltools 0.4.0 2019-10-04 [1] CRAN (R 3.6.0)   
#> htmlwidgets 1.5.1 2019-10-08 [1] CRAN (R 3.6.0)   
#> httr 1.4.1 2019-08-05 [1] CRAN (R 3.6.0)   
#> janeaustenr 0.1.5 2017-06-10 [1] CRAN (R 3.6.0)   
#> jpeg 0.1-8.1 2019-10-24 [1] CRAN (R 3.6.0)   
#> jsonlite 1.6.1 2020-02-02 [1] CRAN (R 3.6.0)   
#> knitr 1.28 2020-02-06 [1] CRAN (R 3.6.0)   
#> labeling 0.3 2014-08-23 [1] CRAN (R 3.6.0)   
#> lattice \* 0.20-40 2020-02-19 [1] CRAN (R 3.6.0)   
#> latticeExtra 0.6-29 2019-12-19 [1] CRAN (R 3.6.0)   
#> lifecycle 0.2.0 2020-03-06 [1] CRAN (R 3.6.0)   
#> lubridate 1.7.4 2018-04-11 [1] CRAN (R 3.6.0)   
#> magrittr 1.5 2014-11-22 [1] CRAN (R 3.6.0)   
#> Matrix 1.2-18 2019-11-27 [1] CRAN (R 3.6.0)   
#> memoise 1.1.0 2017-04-21 [1] CRAN (R 3.6.0)   
#> modelr 0.1.6 2020-02-22 [1] CRAN (R 3.6.0)   
#> modeltools 0.2-23 2020-03-05 [1] CRAN (R 3.6.0)   
#> munsell 0.5.0 2018-06-12 [1] CRAN (R 3.6.0)   
#> nlme 3.1-145 2020-03-04 [1] CRAN (R 3.6.0)   
#> NLP 0.2-0 2018-10-18 [1] CRAN (R 3.6.0)   
#> nnet 7.3-13 2020-02-25 [1] CRAN (R 3.6.0)   
#> pillar 1.4.3 2019-12-20 [1] CRAN (R 3.6.0)   
#> pkgbuild 1.0.6 2019-10-09 [1] CRAN (R 3.6.0)   
#> pkgconfig 2.0.3 2019-09-22 [1] CRAN (R 3.6.0)   
#> pkgload 1.0.2 2018-10-29 [1] CRAN (R 3.6.0)   
#> plyr 1.8.6 2020-03-03 [1] CRAN (R 3.6.0)   
#> png 0.1-7 2013-12-03 [1] CRAN (R 3.6.0)   
#> prettyunits 1.1.1 2020-01-24 [1] CRAN (R 3.6.0)   
#> processx 3.4.2 2020-02-09 [1] CRAN (R 3.6.0)   
#> ps 1.3.2 2020-02-13 [1] CRAN (R 3.6.0)   
#> purrr \* 0.3.3 2019-10-18 [1] CRAN (R 3.6.0)   
#> R6 2.4.1 2019-11-12 [1] CRAN (R 3.6.0)   
#> RColorBrewer 1.1-2 2014-12-07 [1] CRAN (R 3.6.0)   
#> Rcpp 1.0.4 2020-03-17 [1] CRAN (R 3.6.0)   
#> readr \* 1.3.1 2018-12-21 [1] CRAN (R 3.6.0)   
#> readxl \* 1.3.1 2019-03-13 [1] CRAN (R 3.6.0)   
#> remotes 2.1.1 2020-02-15 [1] CRAN (R 3.6.0)   
#> reprex 0.3.0 2019-05-16 [1] CRAN (R 3.6.0)   
#> reshape2 1.4.3 2017-12-11 [1] CRAN (R 3.6.0)   
#> rlang 0.4.5 2020-03-01 [1] CRAN (R 3.6.0)   
#> rmarkdown 2.1 2020-01-20 [1] CRAN (R 3.6.0)   
#> rpart 4.1-15 2019-04-12 [1] CRAN (R 3.6.0)   
#> rprojroot 1.3-2 2018-01-03 [1] CRAN (R 3.6.0)   
#> rstudioapi 0.11 2020-02-07 [1] CRAN (R 3.6.0)   
#> rvest 0.3.5 2019-11-08 [1] CRAN (R 3.6.0)   
#> scales 1.1.0 2019-11-18 [1] CRAN (R 3.6.0)   
#> sessioninfo 1.1.1 2018-11-05 [1] CRAN (R 3.6.0)   
#> slam 0.1-47 2019-12-21 [1] CRAN (R 3.6.0)   
#> SnowballC 0.6.0 2019-01-15 [1] CRAN (R 3.6.0)   
#> stringi 1.4.6 2020-02-17 [1] CRAN (R 3.6.0)   
#> stringr \* 1.4.0 2019-02-10 [1] CRAN (R 3.6.0)   
#> survival \* 3.1-11 2020-03-07 [1] CRAN (R 3.6.0)   
#> testthat 2.3.2 2020-03-02 [1] CRAN (R 3.6.0)   
#> tibble \* 2.1.3 2019-06-06 [1] CRAN (R 3.6.0)   
#> tidyr \* 1.0.2 2020-01-24 [1] CRAN (R 3.6.0)   
#> tidyselect 1.0.0.9000 2020-02-23 [1] Github (r-lib/tidyselect@bb145af)   
#> tidytext \* 0.2.3 2020-03-04 [1] CRAN (R 3.6.0)   
#> tidyverse \* 1.3.0 2019-11-21 [1] CRAN (R 3.6.0)   
#> tm 0.7-7 2019-12-12 [1] CRAN (R 3.6.0)   
#> tokenizers 0.2.1 2018-03-29 [1] CRAN (R 3.6.0)   
#> topicmodels \* 0.2-9 2019-12-03 [1] CRAN (R 3.6.0)   
#> usethis 1.5.1.9000 2020-03-20 [1] Github (r-lib/usethis@4cfa070)   
#> vctrs 0.2.4 2020-03-10 [1] CRAN (R 3.6.0)   
#> withr 2.1.2 2018-03-15 [1] CRAN (R 3.6.0)   
#> xfun 0.12 2020-01-13 [1] CRAN (R 3.6.0)   
#> xml2 1.2.5 2020-03-11 [1] CRAN (R 3.6.0)   
#> yaml 2.2.1 2020-02-01 [1] CRAN (R 3.6.0)   
#>   
#> [1] /Library/Frameworks/R.framework/Versions/3.6/Resources/library

The current Git commit details are: u

#> Local: master /Users/bmarwick/Desktop/archyconfgender  
#> Remote: master @ origin (https://github.com/yichun33/archyconfgender)  
#> Head: [f3f3ad7] 2020-03-23: 18 Title