Teaching experimentation on raw data using multiverse analysis

Nathan Taback

Associate Professor, Teaching Stream

Director, Data Science Programs

Department of Statistical Sciences, University of Toronto

2022-05-12

Talk Outline

1. Introduction
2. Multiverse analysis
3. Teaching Multiverse analysis
4. Available packages - (R: `multiverse`, `mverse`. python: `boba`)
5. Example: effect of fertility on religiosity--using `mverse` to compute a multiverse analysis

HOW DID WE GO FROM STATISTICIAN --> DATA SCIENTIST?

Google (early 2000s)

- Google first learned how to capture surplus behavioral data, more than what they needed for services.

- Google used this data to compute prediction products that they could sell to their business customers - advertisers.

- Behavioral surplus became the defining element of Google's success.

- Hal Varian developed Google's early advertising model that clarified the true nature of Google's business, "All of a sudden, we realized we were in the auction business." (Zuboff, S. 2019. Surveillance Capitalism and the Challenge of Collective Action)

Google (early 2000s)

- The precision of the Carol Brady data was eye-opening for some. 'It was like trying an electron microscope for the first time … It was like a moment-by-moment barometer.'' said Sergey Brin (Jennifer 8 Lee, Nov. 28, 2002, NYT)

- "… Statistical data analysis is much more appropriately associated with the sciences and with the experimental process in general." (Turkey and Wilk, 1966)

SNAPSHOT
Samples of search-engine queries by users around the world are projected on a wall behind Deb Henigson, a receptionist at Google's offices in Mountain View, Calif.

Postcards From Planet Google

What's on the World's Mind? Search Queries Yield a Constant Catalog

By JENNIFER 8 LEE

MOUNTAIN VIEW, Calif.

AT Google's squat headquarters off

island … pickup lines to get women … auto theft fraud how to. …

Stare at Live Query long enough, and you feel that you are watching the collective consciousness of the world stream by.

said Craig Silverstein, director of technology at Google. (To keep Live Query G-rated, Google filters out sex-related searches, though less successfully with foreign languages.)

The hybrid profession of data scientist

"data scientists are … hybrids, who combine generally conflictive roles as both generalists and specialists; technicians and communicators; data exploiters and data ethicists … they [do] … explorative work [to] establish growth for digital capitalism by generating behavioural patterns that allow for personalization, customization and optimization practices."

(Dorschel, 2021)

Data Experimentation

- Tukey referred to "Specific techniques, such as a three-way analysis of variance as "data-analytic hardware".

- The *mystery and art of when and why* to use such techniques are "data-analytic software", which is "very soft indeed".

- One of the most important skills needed for the modern data scientist and statistician is "when and why to use techniques" on "raw data" to find meaning and insight.

MULTIVERSE ANALYSIS

Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

[a]Department of Business Administration and [b]Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and [c]Women and Gender in Global Perspectives, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and [d]Department of Statistics, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287-3701

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes ... violence and destruction (23, 24). We extend these findings to hypothesize that the anticipated severity of a hurricane with ...

- Nine independent coders who were blind to the hypothesis rated the masculinity vs. femininity of historical hurricane names on two items (1 = very masculine, 11 = very feminine, and 1 = very man-like, 11 = very woman-like), which were averaged to compute a masculinity-femininity index (MFI).

- A series of negative binomial regression (Gaussian, Poisson) analyses were performed to investigate effects of perceived masculinity-femininity of hurricane names (MFI) (binary/continuous), minimum pressure, normalized damage (NDAM) (transform), and the interactions among them on the number of deaths (transform) caused by the hurricanes

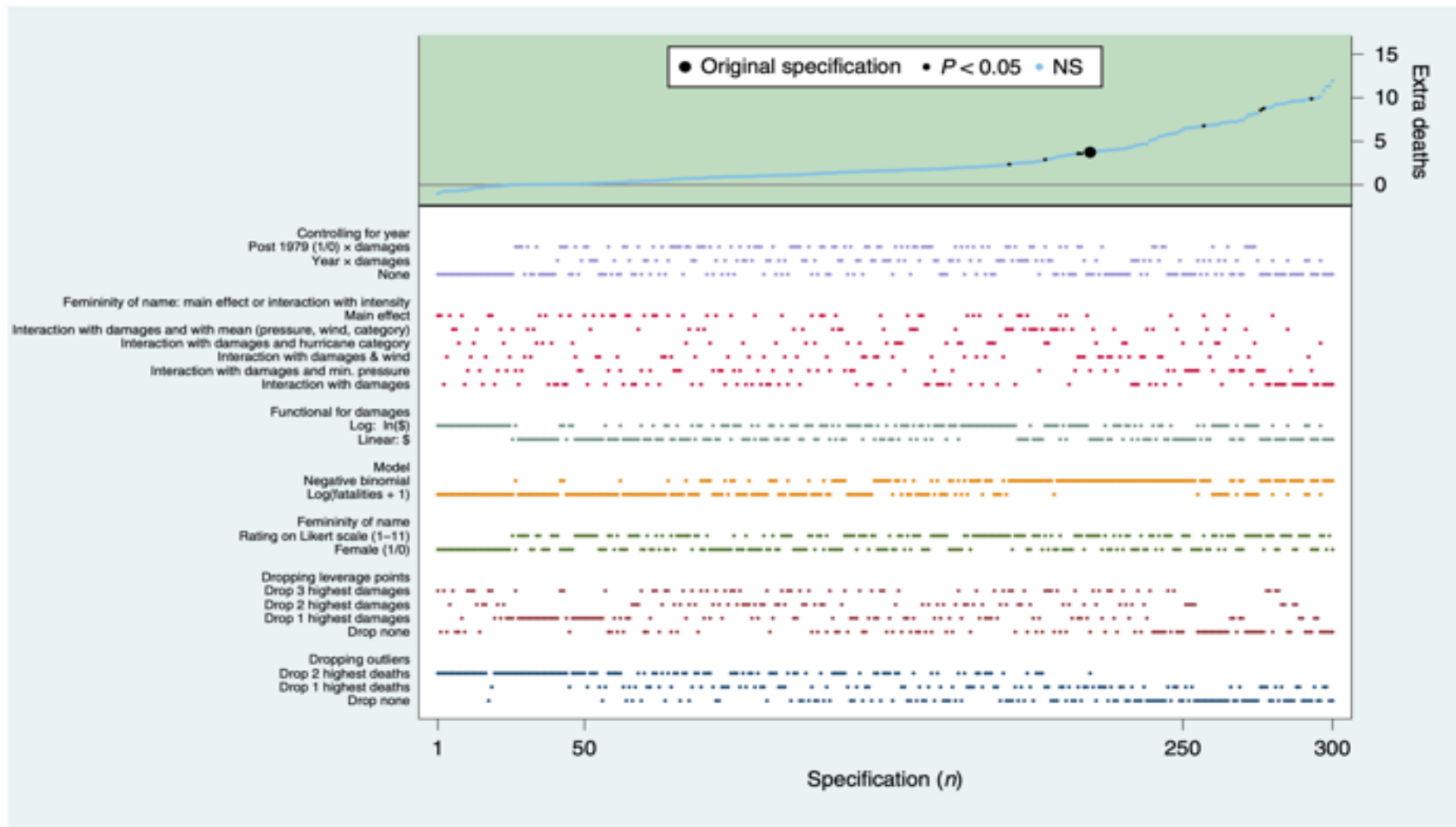Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature human behaviour*, 4(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z
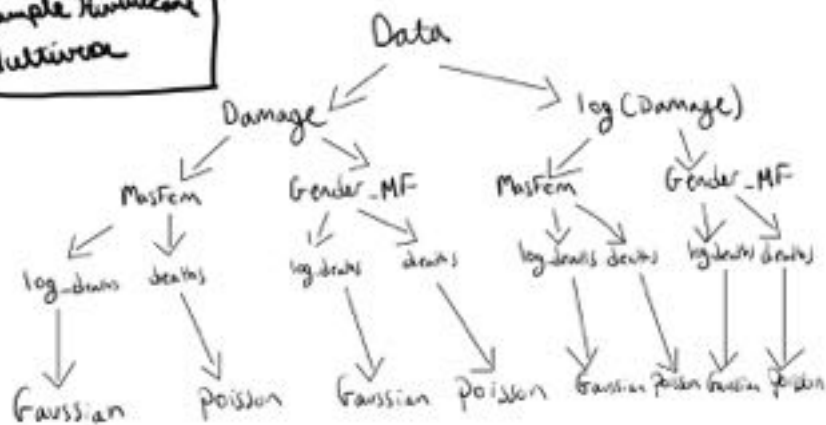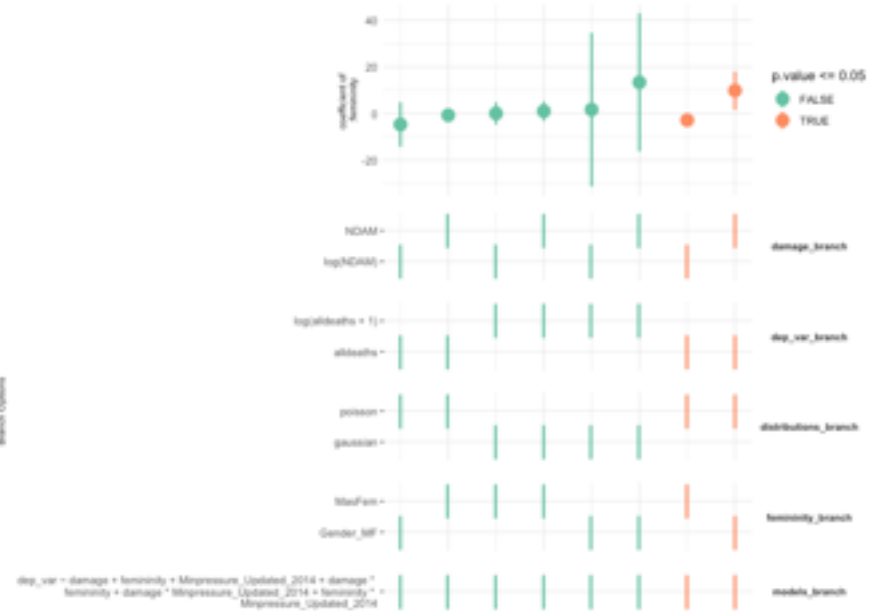
Multiverse Analysis

- "A multiverse analysis starts from the observation that data used in an analysis are usually not just passively recorded in an experiment or an observational study. Rather, data are to a certain extent actively constructed." (Steegen et al. 2016)

- Transformation of variables (e.g., discretization, linear combinations), and creating subsets that may be excluded leads to many possible data sets for analysis.

- "Researchers often select a single (or a few) data processing choices and then present this as the only analysis that ever would have been done. " (Steegen et al. 2016)

Multiverse Analysis

- Choosing among the various possible data sets is often arbitrary.

- A multiverse analysis involves performing the analysis of interest across the whole set of data sets that arise from different ***reasonable*** choices for data processing.

- A multiverse analysis embodies Tukey's idea that data analysis is an iterative process that requires human judgement at every step.

Multiverse Analysis

- *"... our certainty about what will happen in a particular situation does not usually come from directly applicable experiments or theory*, but rather through analogy between situations which are known to behave similarly. Data analysis has ... to be an experimental science ..." (FODA, pg. 63)

- Multiverse analysis provides a paradigm for iterating through judgements made by the analyst, and may be helpful in teaching students about the art of data analysis.

Teaching Multiverse Analysis

- "Data analysis must be iterative to be effective" (Tukey and Wilk, 1966).
- Many (and maybe most others?) teach data analysis by analogy/case studies/etc.

PPDAC as an iterative process
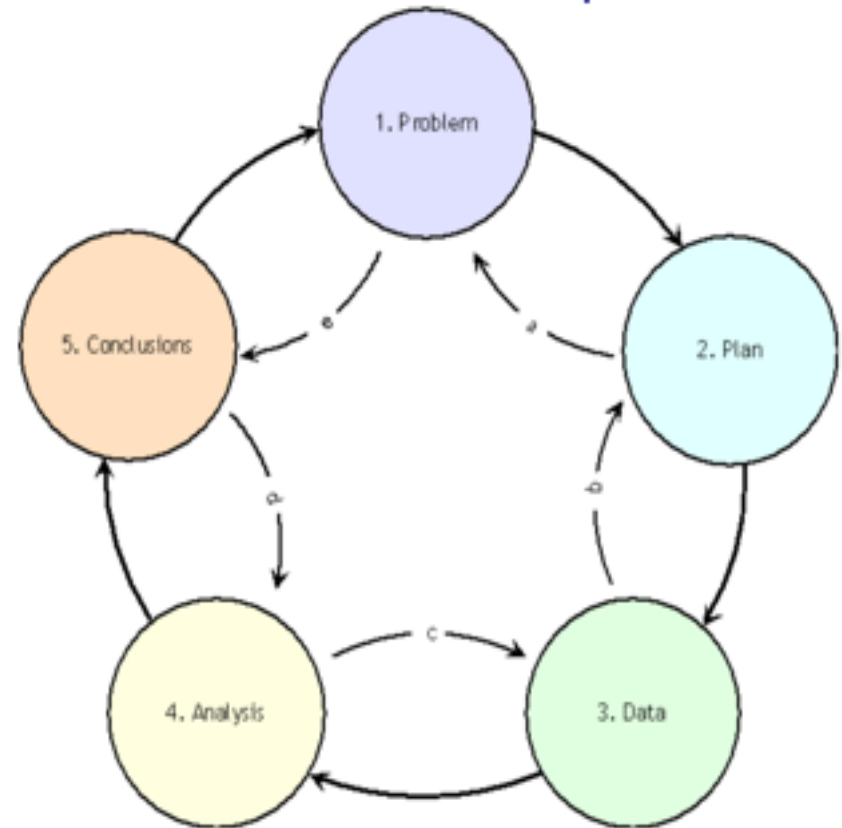
Teaching Multiverse Analysis

- Raw data can be transformed which may change the analysis, results, and conclusions. Yet, these transformations lead to reasonable (and justifiable) choices.

- In other words that there are multiple ways to analyze a data set, leading to, for example, multiple p-values for a parameter.

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time[*]

Andrew Gelman[†] and Eric Loken[‡]

14 Nov 2013

Where Could A Multiverse Analysis Be Incorporated into a Course?

- Course research project involving data analysis
- Learning objectives related to data analysis concepts such as transformation, interpretation, and communication.
- Learning objectives related to computing with data (this is one of the reasons that computing is so important in statistics)

What's Required to Teach Multiverse in a Course?

- A dataset that has dependent and independent variables that can be transformed, or it's of interest to analyse subsets (e.g., data exclusion).

- Computing should be integrated into the course.

Multiverse packages (currently available that I'm aware of ...)

- R: `multiverse`, `mverse`
- python: `boba`
- These packages offer a computational paradigm for the iterative and subjective parts of data analysis.

WHY DEVELOP A NEW PACKAGE TO CONDUCT A MULTIVERSE ANALYSIS?

tidyverse

```r
set.seed(21); mydat <- dplyr::tibble(y = rnorm(100), x = y + runif(100))

mod1 <- stats::lm(y ~ x, data = mydat)

mod2 <- stats::lm(y ~ log(abs(x+1)), data = mydat)

summary1 <- mod1 |> broom::tidy()

summary2 <- mod2 |> broom::tidy()

cbind(universe = c(1, 1, 2, 2), rbind(summary1, summary2))
```

Redundant code

```
##   universe            term     estimate  std.error  statistic      p.value
## 1        1     (Intercept)  -0.46901356 0.03347050 -14.012746 3.856752e-25
## 2        1               x   0.97117012 0.02903206  33.451648 2.059139e-55
## 3        2     (Intercept)  -0.08146011 0.05844837  -1.393711 1.665576e-01
## 4        2 log(abs(x + 1))   0.87853965 0.05952913  14.758146 1.209627e-26
```

multiverse

```r
set.seed(21); mydat <- dplyr::tibble(y = rnorm(100), x = y + runif(100))

M <- multiverse::multiverse()
multiverse::inside(M, {

  mydat <- mydat |> dplyr::mutate(x_branches = branch(x_branch, "none" ~ x, "logx"
~ log(abs(x+1))))

  mod <- stats::lm(y ~ x_branches, data = mydat)

  mod_sum <- mod |> broom::tidy(conf.int = TRUE)
})
 multiverse::expand(M) |> multiverse::extract_variables(mod_sum) |> tidyr::unnest(cols =
 c(mod_sum))
```

Intermediate/Advanced R concepts

```
## # A tibble: 4 × 12
##   .universe x_branch .parameter_assignment .code         .results term    estimate
##       <int> <chr>    <list>                <list>        <list>   <chr>        <dbl>
## 1         1 none     <named list [1]>      <named list> <env>     (Inte… -0.469
## 2         1 none     <named list [1]>      <named list> <env>     x_bra…  0.971
## 3         2 logx     <named list [1]>      <named list> <env>     (Inte… -0.0815
## 4         2 logx     <named list [1]>      <named list> <env>     x_bra…  0.879
## # … with 5 more variables: std.error <dbl>, statistic <dbl>, p.value <dbl>,
## #   conf.low <dbl>, conf.high <dbl>
```

multiverse

```{r}
set.seed(21); mydat <- dplyr::tibble(y = rnorm(100), x = y + runif(100))

M <- multiverse::multiverse()
```

```{multiverse mymverse, inside = M}
mydat <- mydat |> dplyr::mutate(x_branches = branch(x_branch, "none" ~ x, "logx" ~ log(abs(x+1))))

mod <- stats::lm(y ~ x_branches, data = mydat)

mod_sum <- mod |> broom::tidy(conf.int = TRUE)
```

```{r}
multiverse::execute_multiverse(M)

multiverse::expand(M) |> multiverse::extract_variables(mod_sum) |> tidyr::unnest(cols = c(mod_sum))
```

A tibble: 4 × 12

.universe <int>	x_branch <chr>	.parameter_assignment <list>	.code <list>	.results <list>	term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>
1	none	<list [1]>	<list [1]>	<env>	(Intercept)	−0.46901356	0.03347050	−14.012746
1	none	<list [1]>	<list [1]>	<env>	x_branches	0.97117012	0.02903206	33.451648
2	logx	<list [1]>	<list [1]>	<env>	(Intercept)	−0.08146011	0.05844837	−1.393711
2	logx	<list [1]>	<list [1]>	<env>	x_branches	0.87853965	0.05952913	14.758146

4 rows | 1–9 of 12 columns

mverse

```r
set.seed(21); mydat <- dplyr::tibble(y = rnorm(100), x = y + runif(100))

M <- mverse::create_multiverse(mydat)

x_branches <- mverse::mutate_branch(x, log(abs(x + 1)))

M <- M |> mverse::add_mutate_branch(x_branches)

form1 <- mverse::formula_branch(y ~ x_branches)

M <- M |> mverse::add_formula_branch(form1)

lm_mverse(M)

summary(M)
## # A tibble: 4 × 10
##   universe x_branches_branch form1_branch    term        estimate std.error statistic
##   <fct>    <fct>             <fct>           <chr>          <dbl>      <dbl>     <dbl>
## 1 1        x                 y ~ x_branches (Inter…       -0.469     0.0335     -14.0
## 2 1        x                 y ~ x_branches x_bran…        0.971     0.0290      33.5
## 3 2        log(abs(x + 1))   y ~ x_branches (Inter…       -0.0815    0.0584      -1.39
## 4 2        log(abs(x + 1))   y ~ x_branches x_bran…        0.879     0.0595      14.8
## # … with 3 more variables: p.value <dbl>, conf.low <dbl>, conf.high <dbl>
```

Familiar introductory `dplyr` verbs, base R syntax, familiar `tibble` as results summary

`mverse`

- Built as an easier-to-use version of `multiverse`.
- Wrappers for `t.test()`, `lm()`, `glm()`, and `glm.nb()`.
- Don't have to declare a multiverse code-chunk or a multiverse environment using `multiverse::inside()`

Effect of Fertility on Religiosity and Political Attitudes

- Durante et al. (2013) studied the effect of fertility on religiosity and political attitudes.

- Women were classified into high (7 - 14 days) and low (17 - 25 days) fertility groups based on their cycle day at time of survey.

The Menstrual Cycle
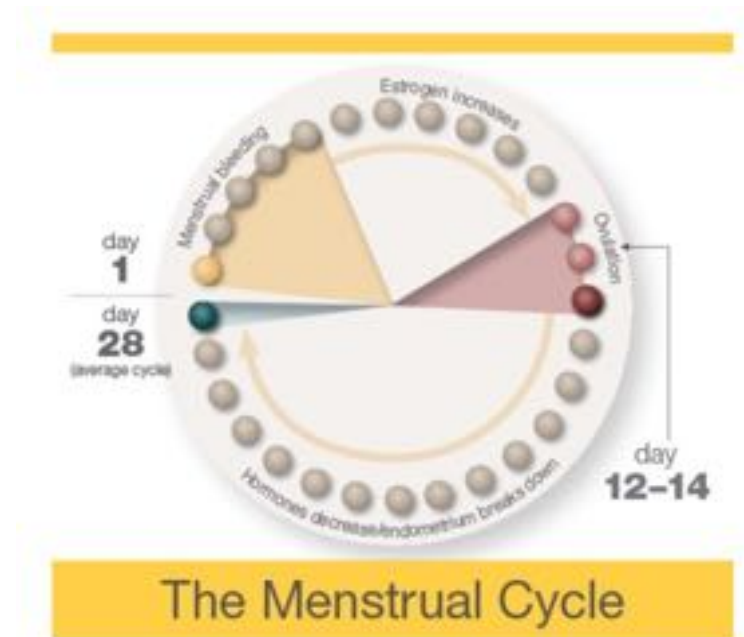
Effect of Fertility on Religiosity and Political Attitudes

```
ComputedCycleLength =
StartDateofLastPeriod –
StartDateofPeriodBeforeLast

NextMenstrualOnset = StartDateofLastPeriod +
ComputedCycleLength

CycleDay = 28 - (NextMenstrualOnset -
DateTesting)
```

Effect of Fertility on Religiosity and Political Attitudes

- Instead of `ComputedCycleLength` we could have used `ReportedCycleLength` or estimated `NextMenstrualOnset = StartDeateNext`.
- This would give us different cycle days and might lead to different women being classified as high/low fertility.

Analysis using `tidyverse`

Data for analysis

```r
four_universe <- durante %>%
    mutate(ComputedCycleLength = StartDateofLastPeriod - StartDateofPeriodBeforeLast,

           NextMenstrualOnset1 = StartDateofLastPeriod + ComputedCycleLength,
           NextMenstrualOnset2 = StartDateofLastPeriod + ReportedCycleLength,

           CycleDay1 = 28 - (NextMenstrualOnset1 - DateTesting),
           CycleDay2 = 28 - (NextMenstrualOnset2 - DateTesting),

           Fertility1 = factor(ifelse(CycleDay1 >= 7 & CycleDay1 <= 14, "high",
                                      ifelse(CycleDay1 >= 17 & CycleDay1 <= 25, "low", NA))),
           Fertility2 = factor(ifelse(CycleDay2 >= 7 & CycleDay2 <= 14, "high",
                                      ifelse(CycleDay2 >= 17 & CycleDay2 <= 25, "low", NA))),

           Relationship1 = factor(ifelse(Relationship==1 | Relationship==2, "Single", "Relationship")),
           Relationship2 = factor(ifelse(Relationship==1, "Single", "Relationship")),

           RelComp = round((Rel1 + Rel2 + Rel3)/3, 2))
```

Modelling

```r
anova(lm(RelComp ~ Fertility1*Relationship1, data = four_universe))
anova(lm(RelComp ~ Fertility1*Relationship2, data = four_universe))
anova(lm(RelComp ~ Fertility2*Relationship1, data = four_universe))
anova(lm(RelComp ~ Fertility2*Relationship2, data = four_universe))
```

Analysis using `tidyverse`

```r
m1 <- lm(RelComp ~ Fertility1*Relationship1, data = four_universe) |> broom::tidy(conf.int=T) |>
select(estimate, p.value, conf.low, conf.high) |> slice(4)

m2 <- lm(RelComp ~ Fertility1*Relationship2, data = four_universe) |> broom::tidy(conf.int=T) |>
select(estimate, p.value, conf.low, conf.high) |> slice(4)

m3 <- lm(RelComp ~ Fertility2*Relationship1, data = four_universe) |> broom::tidy(conf.int=T) |>
select(estimate, p.value, conf.low, conf.high) |> slice(4)

m4 <- lm(RelComp ~ Fertility2*Relationship2, data = four_universe) |> broom::tidy(conf.int=T) |>
select(estimate, p.value, conf.low, conf.high) |> slice(4)

cbind(models = c("Fert1:Rel1", "Fert1:Rel2", "Fert2:Rel1", "Fert2:Rel2"),rbind(m1,m2,m3,m4)) |>
  ggplot(aes(models, estimate, color = ifelse(p.value<=0.01, "True", "False"))) +
geom_pointrange(aes(ymin=conf.low, ymax=conf.high)) +
  labs(color = "p-value <= 0.01") + ylab("Coefficiant of Fertility & Relationship") + ggtitle("99% CI of
Fertility & Relationship")
```

Redundant code

Use `mverse` to Define Different Options

Data preparation

```r
library(mverse)
library(tidyverse)

durante_mv_dat <- durante %>% dplyr::mutate(ComputedCycleLength = StartDateofLastPeriod -
                                                                  StartDateofPeriodBeforeLast,
                           NextMenstrualOnset1 = StartDateofLastPeriod + ComputedCycleLength,
                           NextMenstrualOnset2 = StartDateofLastPeriod + ReportedCycleLength,
                           CycleDay1 = 28 - (NextMenstrualOnset1 - DateTesting),
                           CycleDay2 = 28 - (NextMenstrualOnset2 - DateTesting))
```

`mverse`

```r
durante_mv <- mverse::create_multiverse(durante_mv_dat)

Fert_br <- mverse::mutate_branch(Fertility1 = factor(ifelse(CycleDay1 >= 7 & CycleDay1 <= 14, "high",
                                               ifelse(CycleDay1 >= 17 & CycleDay1 <= 25, "low", NA))),
                    Fertility2 = factor(ifelse(CycleDay2 >= 7 & CycleDay2 <= 14, "high",
                                               ifelse(CycleDay2 >= 17 & CycleDay2 <= 25, "low", NA))))

Relationship_br <- mverse::mutate_branch(Relationship1 = factor(ifelse(Relationship==1 |
                                                  Relationship==2,"Single","Relationship")),
                        Relationship2 = factor(ifelse(Relationship==1, "Single", "Relationship")))

Relig_br <- mverse::mutate_branch(RelComp = round((Rel1 + Rel2 + Rel3)/3, 2))

durante_mv <- durante_mv %>% mverse::add_mutate_branch(Fert_br, Relationship_br, Relig_br)

formula_br <- mverse::formula_branch(Relig_br ~ Fert_br*Relationship_br)

durante_mv <- durante_mv %>% mverse::add_formula_branch(formula_br)
```

`mutate_branch()` defines options for fertility and relationship computations

summary.lm_mverse()

```
mverse::lm_mverse(durante_mv)

summary(durante_mv) |> select(universe, term, estimate, std.error, statistic, p.value, conf.low,
conf.high)
```
```
A tibble: 16 × 8
universe term   estimate std.error statistic p.value conf.low conf.high
<fct>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 1   (Intercept)   6.37   0.404   15.8 4.21e-41   5.58   7.17
2 1   Fert_brlow   -1.20   0.535   -2.24 2.57e- 2   -2.25   -0.147
3 1   Relationsh…   -1.46   0.540   -2.70 7.34e- 3   -2.52   -0.395
4 1   Fert_brlow…   2.03   0.716   2.84 4.89e- 3   0.621   3.44
5 2   (Intercept)   5.84   0.323   18.1 9.35e-50   5.20   6.47
6 2   Fert_brlow   -0.660   0.429   -1.54 1.25e- 1   -1.50   0.185
7 2   Relationsh…   -0.912   0.582   -1.57 1.18e- 1   -2.06   0.234
8 2   Fert_brlow…   1.93   0.772   2.50 1.30e- 2   0.410   3.45
9 3   (Intercept)   6.55   0.412   15.9 1.06e-39   5.74   7.37
10 3   Fert_brlow   -1.33   0.566   -2.34 1.99e- 2   -2.44   -0.211
11 3   Relationsh…   -1.73   0.547   -3.17 1.70e- 3   -2.81   -0.658
12 3   Fert_brlow…   2.26   0.761   2.96 3.35e- 3   0.756   3.75
13 4   (Intercept)   5.81   0.325   17.9 1.89e-46   5.17   6.46
14 4   Fert_brlow   -0.413   0.454   -0.910 3.64e- 1   -1.31   0.481
15 4   Relationsh…   -0.871   0.610   -1.43 1.55e- 1   -2.07   0.330
16 4   Fert_brlow…   1.25   0.855   1.46 1.46e- 1   -0.437   2.93
```
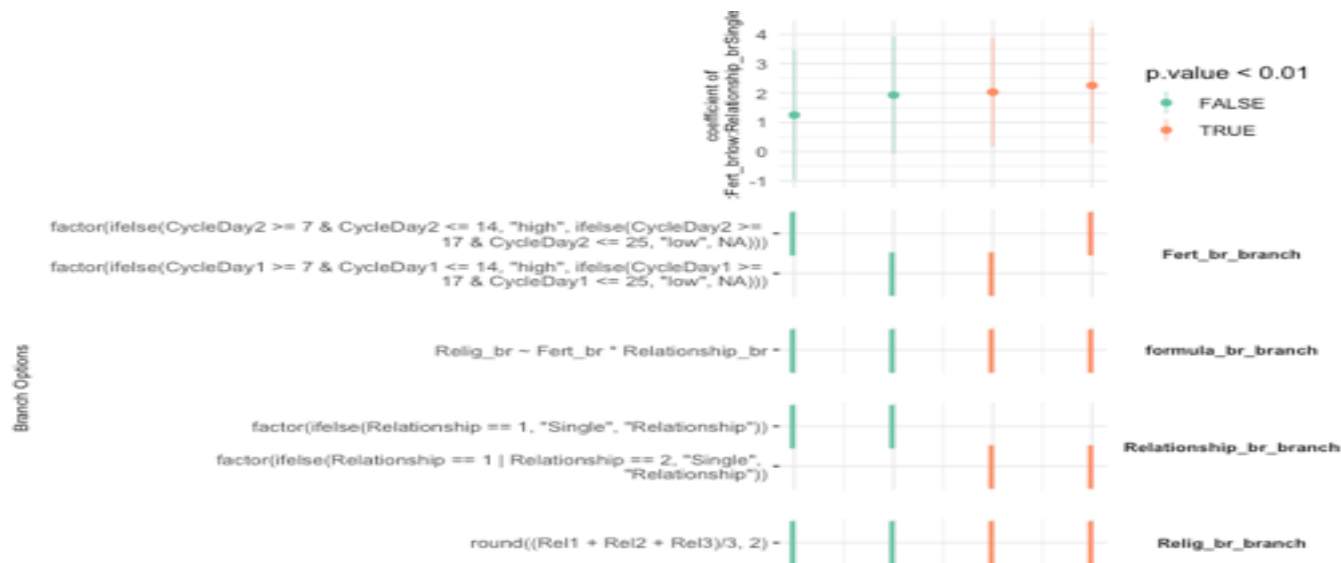
# multiverse_tree

```
mverse::multiverse_tree(durante_mv, label = TRUE, branches =
c("Relationship_br", "Fert_br"), label_size = 2, label_angle = 5)
```
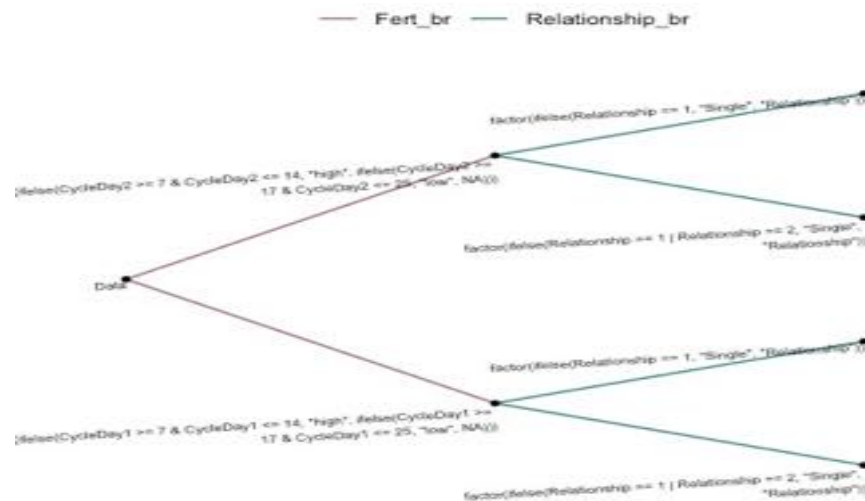
spec_curve()

```
mverse::spec_curve(durante_mv, var =
"Fert_brlow:Relationship_brSingle", color = p.value < 0.01,
grid_size = 10, conf.level = 0.99)
```

What have we learned?

- There have been major shifts in recommended statistics curriculum over the past decade to include computing with data in the practice of statistics (e.g., Nolan and Temple Lang, 2010).

- Modern statistics/data science students should gain experience with the "mystery and art" of data analysis.

- `mverse` can be viewed as a *computational paradigm* for gaining experience with the iterative and subjective parts of data analysis.

This work would not have been possible without …

---

multiverse: 'Explorable Multiverse' Data Analysis and Reports

Implement 'multiverse' style analyses (Steegen S., Tuerlinckx F, Gelman A., Vanpaemel, W., 2016) <doi:10.1177/1745691616658637>, (Dragicevic P., Jansen Y., Sarma A., Kay M., Chevalier F., 2019) <doi:10.1145/3290605.3300295> to show the robustness of statistical inference. 'Multiverse analysis' is a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are. The 'multiverse' package (Sarma A., Kale A., Moon M., Taback N., Chevalier F., Hullman J., Kay M., 2021) <doi:10.31219/osf.io/yfbwm> allows users to concisely and flexibly implement 'multiverse-style' analysis, which involve declaring alternate ways of performing an analysis step, in R and R Notebooks.

Version:	0.5.0
Depends:	R (≥ 3.5.0), knitr (≥ 1.3.2)
Imports:	dplyr (≥ 0.8.1), purrr (≥ 0.3.2), rlang (≥ 0.4.0), R6, methods, tidyr (≥ 1.0.0), tibble, magrittr, tidyselect, formatR, collections, evaluate, rstudioapi, berryFunctions
Suggests:	ggplot2 (≥ 3.0.0), testthat (≥ 2.1.0), highr, lubridate, rmarkdown, covr, broom, boot, gganimate, gifski, forcats, stringr, cowplot, tidybayes, png, stringi, modelr, styler
Published:	2021-06-01
Author:	Abhraneel Sarma [aut, cre], Michael Moon [aut], Matthew Kay [aut], Alex Kale [ctb], Nathan Taback [ctb], Fanny Chevalier [ctb], Jessica Hullman [ctb], Pierre Dragicevic [ctb], Yvonne Jansen [ctb]
Maintainer:	Abhraneel Sarma <abhraneel at u.northwestern.edu>

---

mverse: Tidy Multiverse Analysis Made Simple

Extends 'multiverse' package (Sarma A., Kale A., Moon M., Taback N., Chevalier F., Hullman J., Kay M., 2021) <doi:10.31219/osf.io/yfbwm>, which allows users perform to create explorable multiverse analysis in R. This extension provides an additional level of abstraction to the 'multiverse' package with the aim of creating user friendly syntax to researchers, educators, and students in statistics. The 'mverse' syntax is designed to allow piping and takes hints from the 'tidyverse' grammar. The package allows users to define and inspect multiverse analysis using familiar syntax in R.

Version:	0.1.0
Depends:	R (≥ 3.6), multiverse
Imports:	magrittr (≥ 1.5), rlang, dplyr, tidyr, tidyselect, stringr, stats, broom, grDevices, igraph, ggraph, ggplot2, cowplot
Suggests:	tidyverse, scales, MASS, testthat (≥ 2.1.0), pkgdown (≥ 1.5.1), covr, knitr, rmarkdown, kableExtra
Published:	2022-05-23
Author:	Michael Jongho Moon [aut, cre], Haoda Li [aut], Mingwei Xu [aut], Nathan Taback [aut], Fanny Chevalier [aut], Alison Gibbs [ctb]
Maintainer:	Michael Jongho Moon <michael.moon at mail.utoronto.ca>