MIAMI UNIVERSITY

Lady Tasting Tea Lineups for Visual Inference

Karsten Maurer

July 31, 2019

Lady Tasting Tea [1]

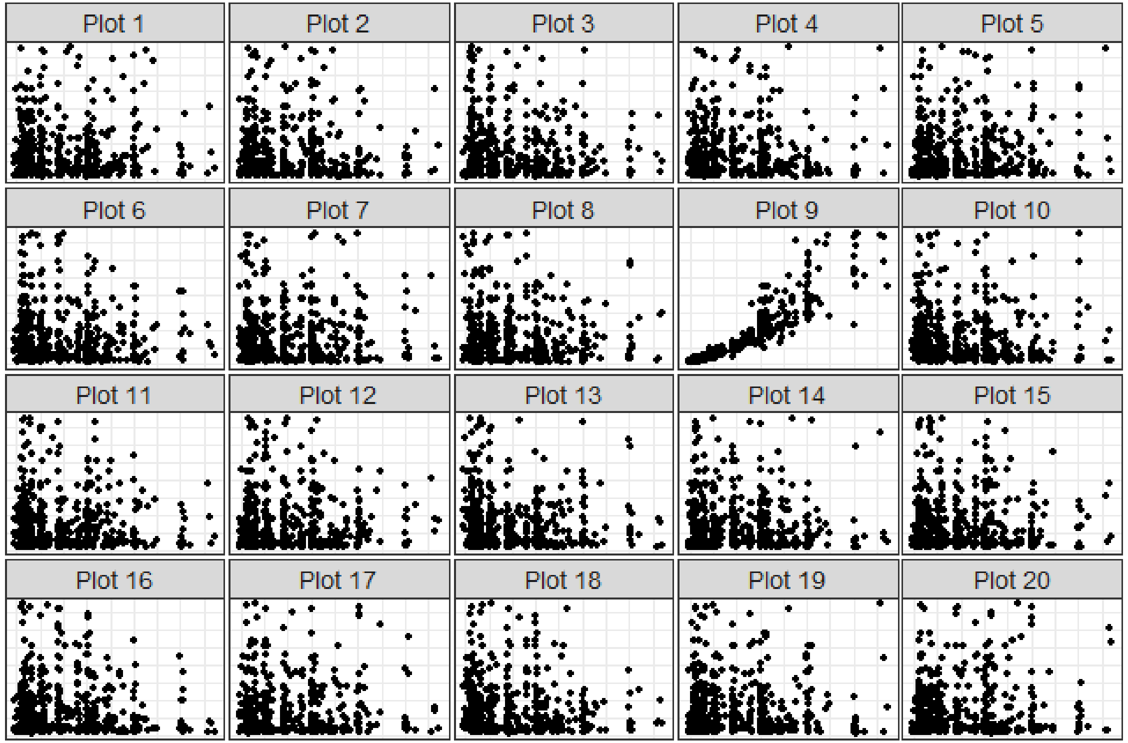
- Ronald Fisher and Muriel Bristol working at Rothamsted
- Claimed she could tell if tea or milk added first
- Blind taste test: four milk first, four tea first
- Fisher's Exact Test $\rightarrow$ follows hypergeometric if guessing

[1] Fisher, R. A. (1960)

Lineups

Can the witness pick the criminal out of a randomized lineup of people?

Lineups (for Visual Inference) [2]



Can the statistician pick the real data out of a randomized lineup of data simulated under the null?

[2] Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., & Wickham, H. (2009)

Lineups (for Visual Inference)

- Plot of real data hidden in set of K-1 plots of data generated under null model

Lineup Interpretations - One Viewer

- If a viewer can pick out the real data (p-value=1/K) $\rightarrow$ reject the null
- If a viewer can't pick out the real data (p-value=1) $\rightarrow$ fail to reject the null

Lineup Interpretations - Many Viewers

- Each person attempts to pick out real data
- Sum of correct guesses distributed Binomial(n,1/K) if viewers independent

Our Work

- Research Team: George Woodbury, Seonjin Kim and myself

- Work started with George's masters project improving visual inference with one person

- After masters, George came back with a new idea...

--

Mashup: Tea Tasting + Lineup Plots

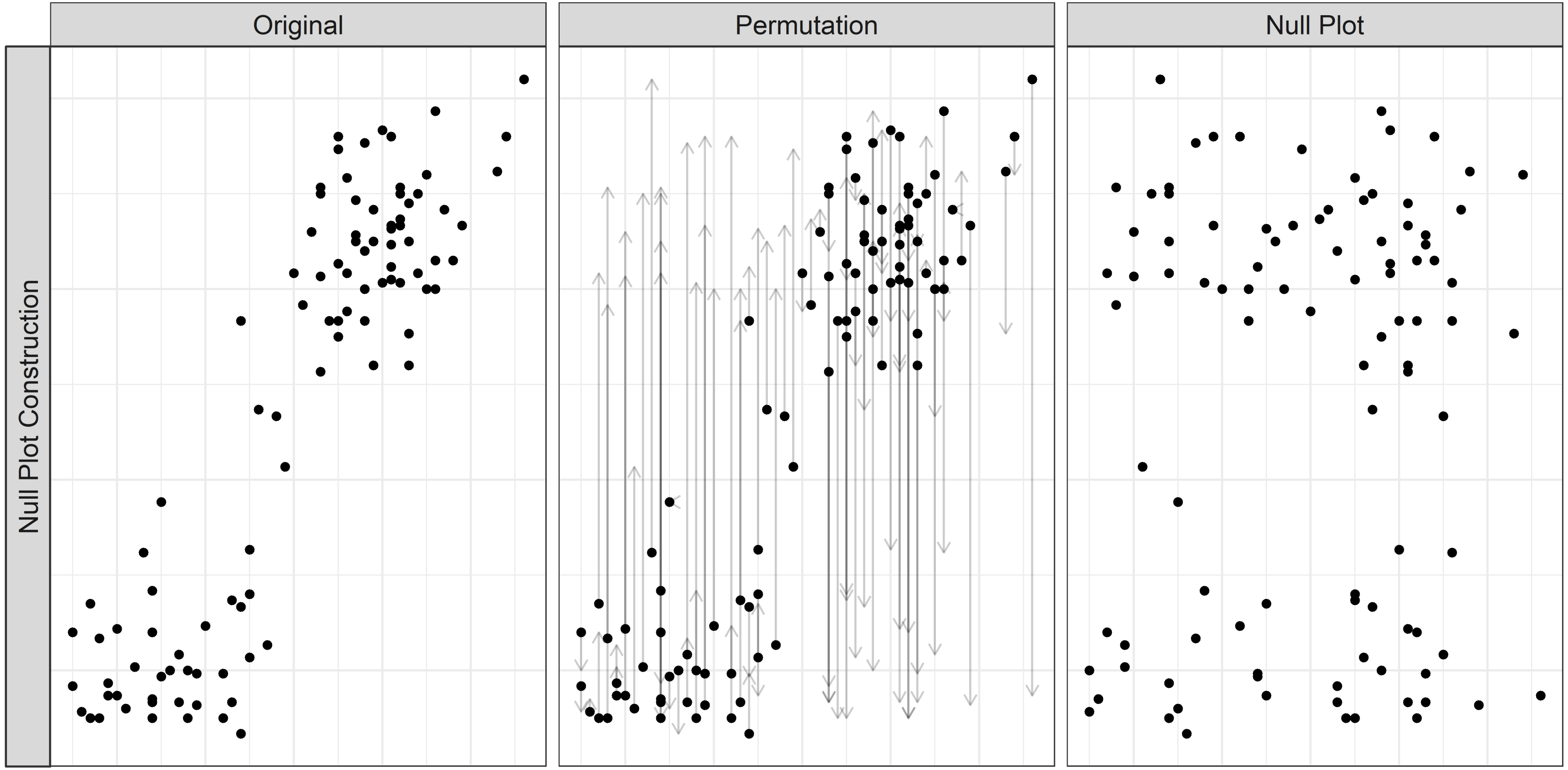
- Applying experimental design from the lady tasting tea

- Randomized lineup of with multiple null and *multiple* target plots

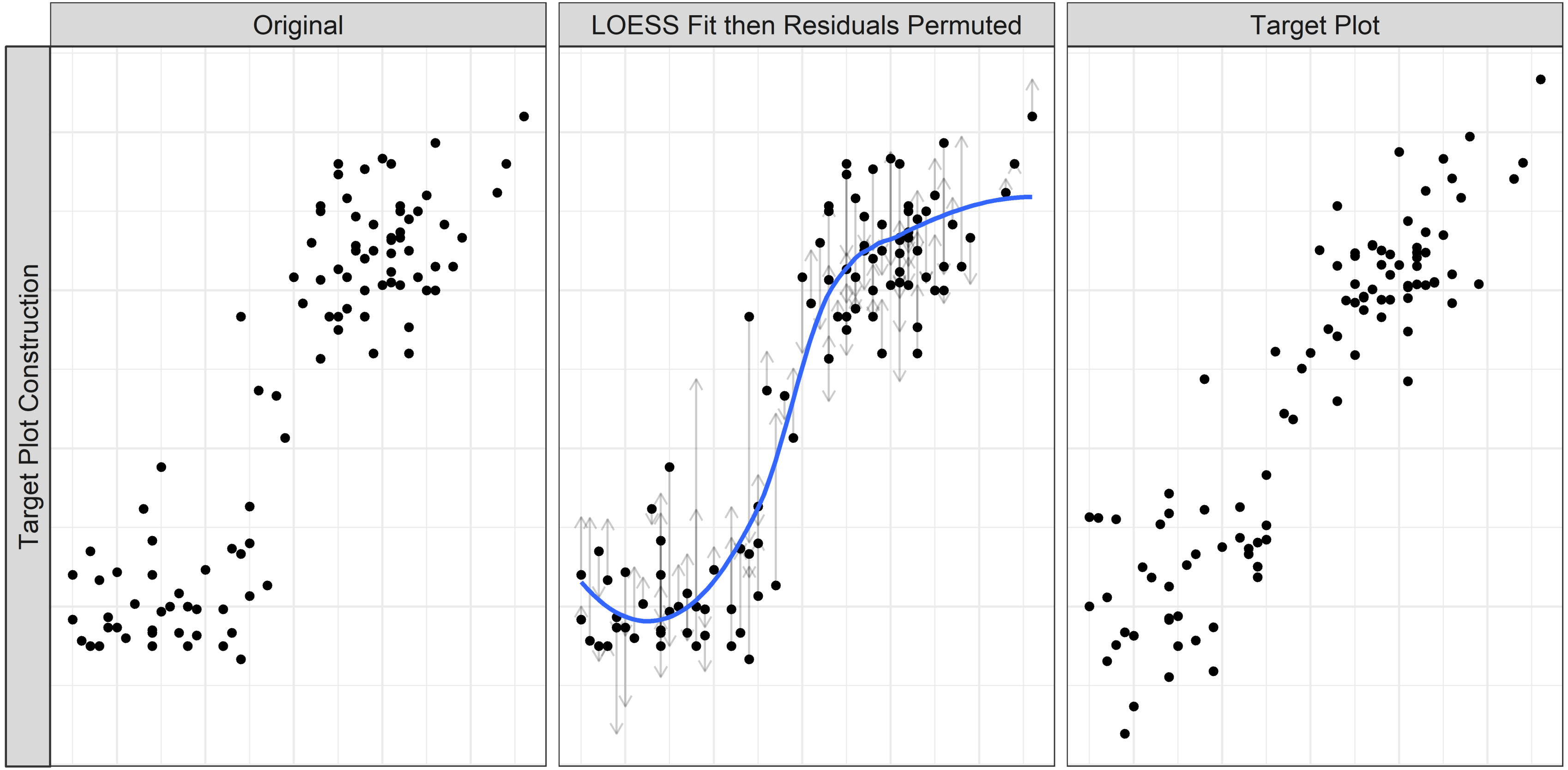
- Viewer (*tea-taster*) tasked with identifying target plots

Methods

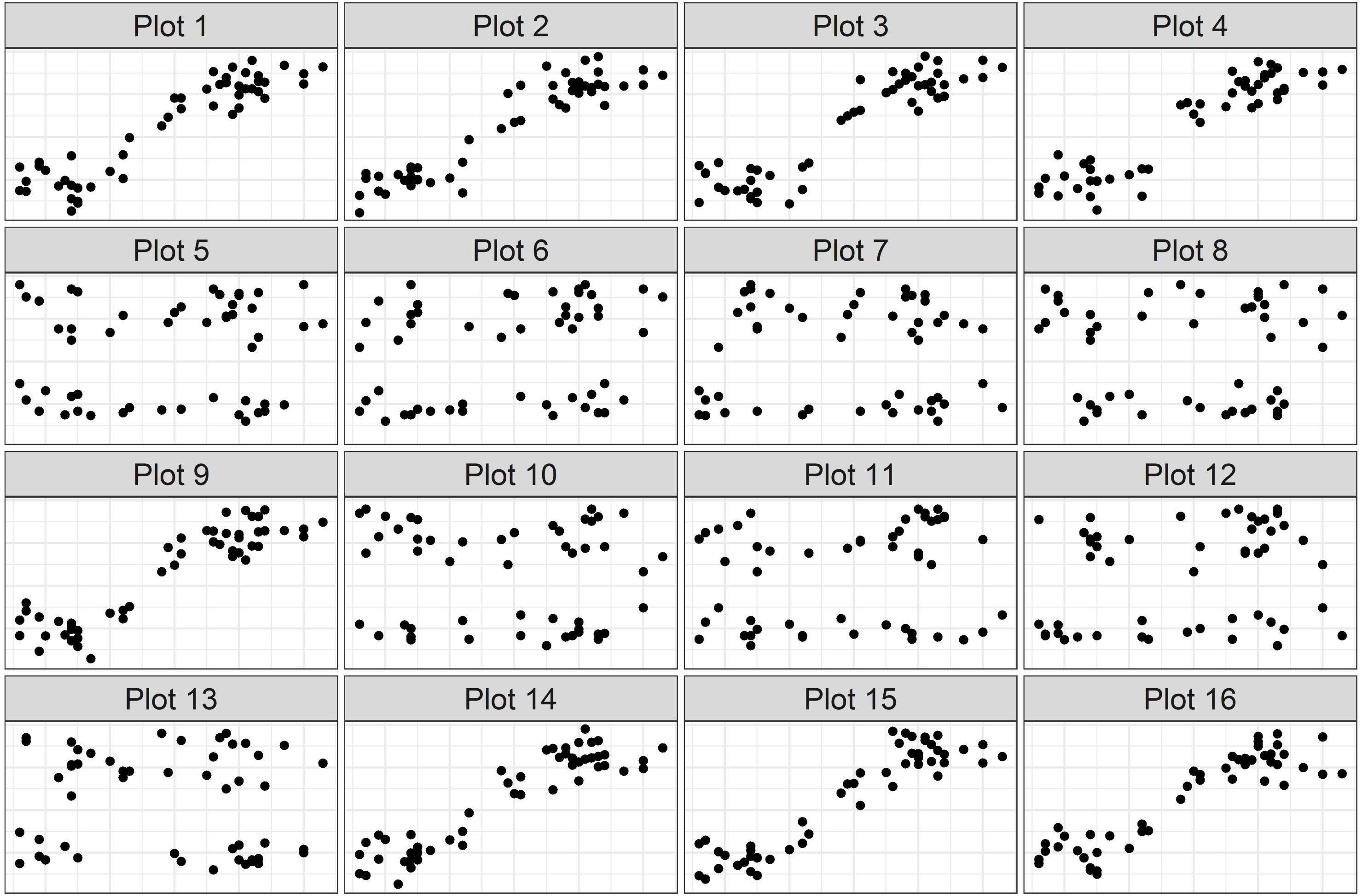
Generating Null Plots for Tea-Tasting Lineups



Generating Target Plots for Tea-Tasting Lineups



Example Tea Tasting Lineup



MIAMI UNIVERSITY

Properties

- Looks similar to traditional lineup for visual inference but task different
- p-values for single TT lineup evaluation are no longer binary
- p-values follow hypergeometric *if null and target plots indistinguishable for data generated by the null*

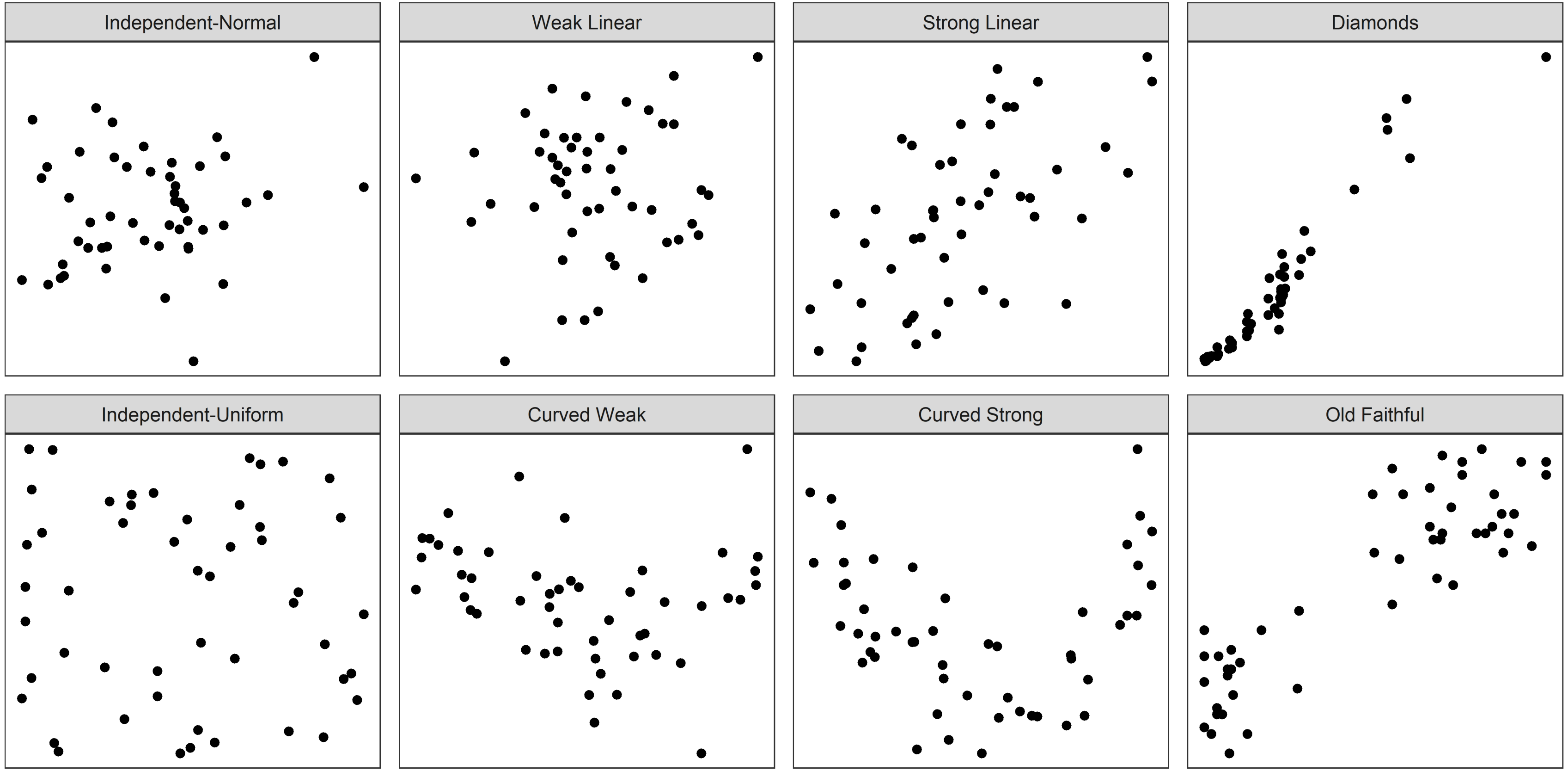
Open Questions

1. Does the TT lineup have discriminative power in practice?

2. Are correct guess counts hypergeometrically distributed for a true null?

Survey

- Participants presented with eight TT lineups based on eight datasets



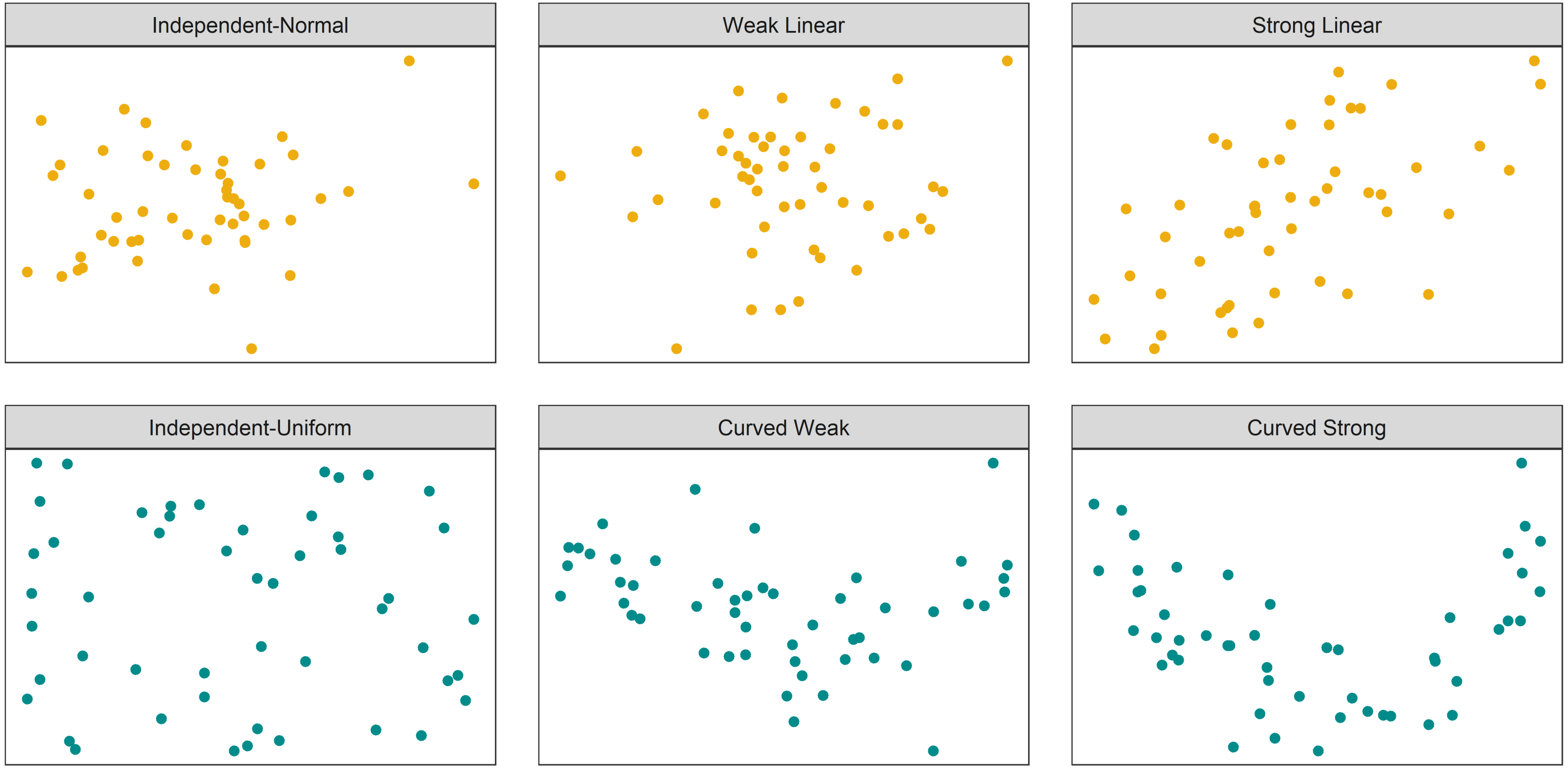
MIAMI UNIVERSITY

Survey

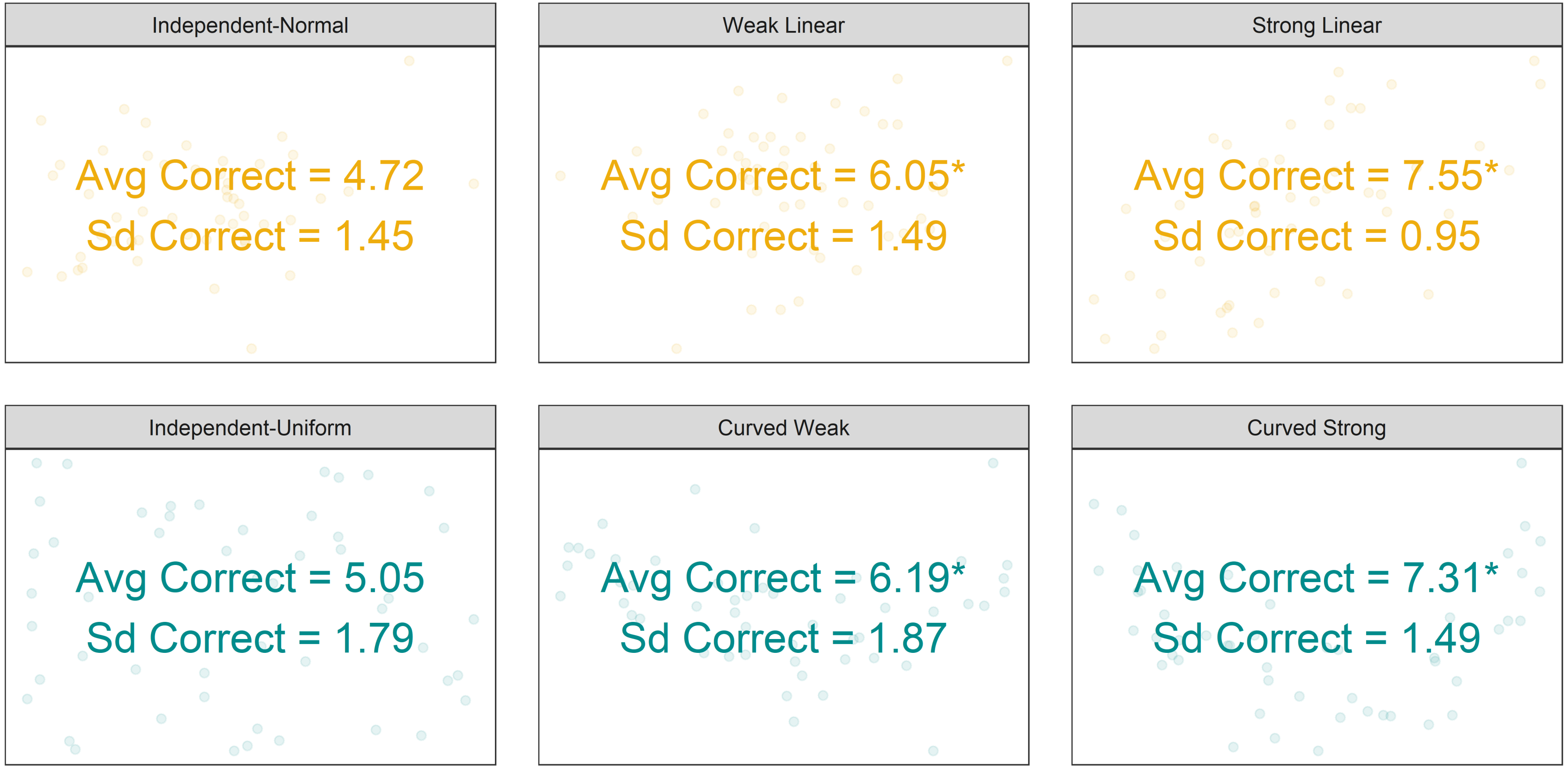
- Administration:
  - distributed to students and faculty of Miami University Stat Department
  - Anonymous drop-box for return
- Content
  - Tea-tasting lineups from 8 different datasets
  - Each lineup had 8 null plots and 8 target plots
- Randomization:
  - Datasets uniquely simulated/sampled for each survey
  - Random order of lineups on survey
  - Random plot ordering within lineups
  - Permutation step for each plot construction
- 45 participants
  - 24 Undergrad, 14 Grad, 6 Faculty, 1 Unknown
  - 41 completed all eight lineups

Results

1. Does the TT lineup have discriminative power in practice?

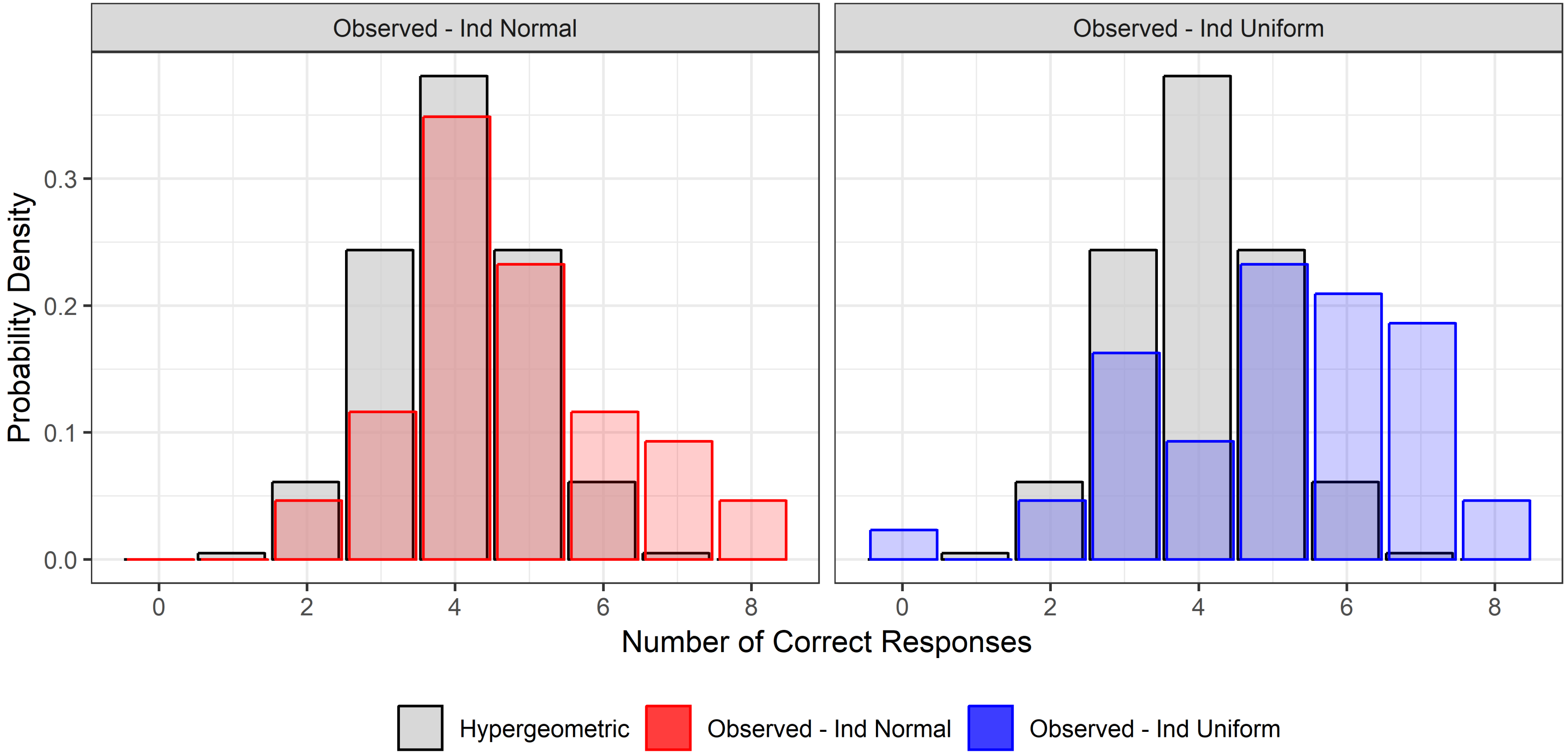


1. Does the TT lineup have discriminative power in practice? Yes!



* t-test strongly suggests more correct answers than lower strength relationship

2. Are the TT lineup p-values hypergeometrically distributed?



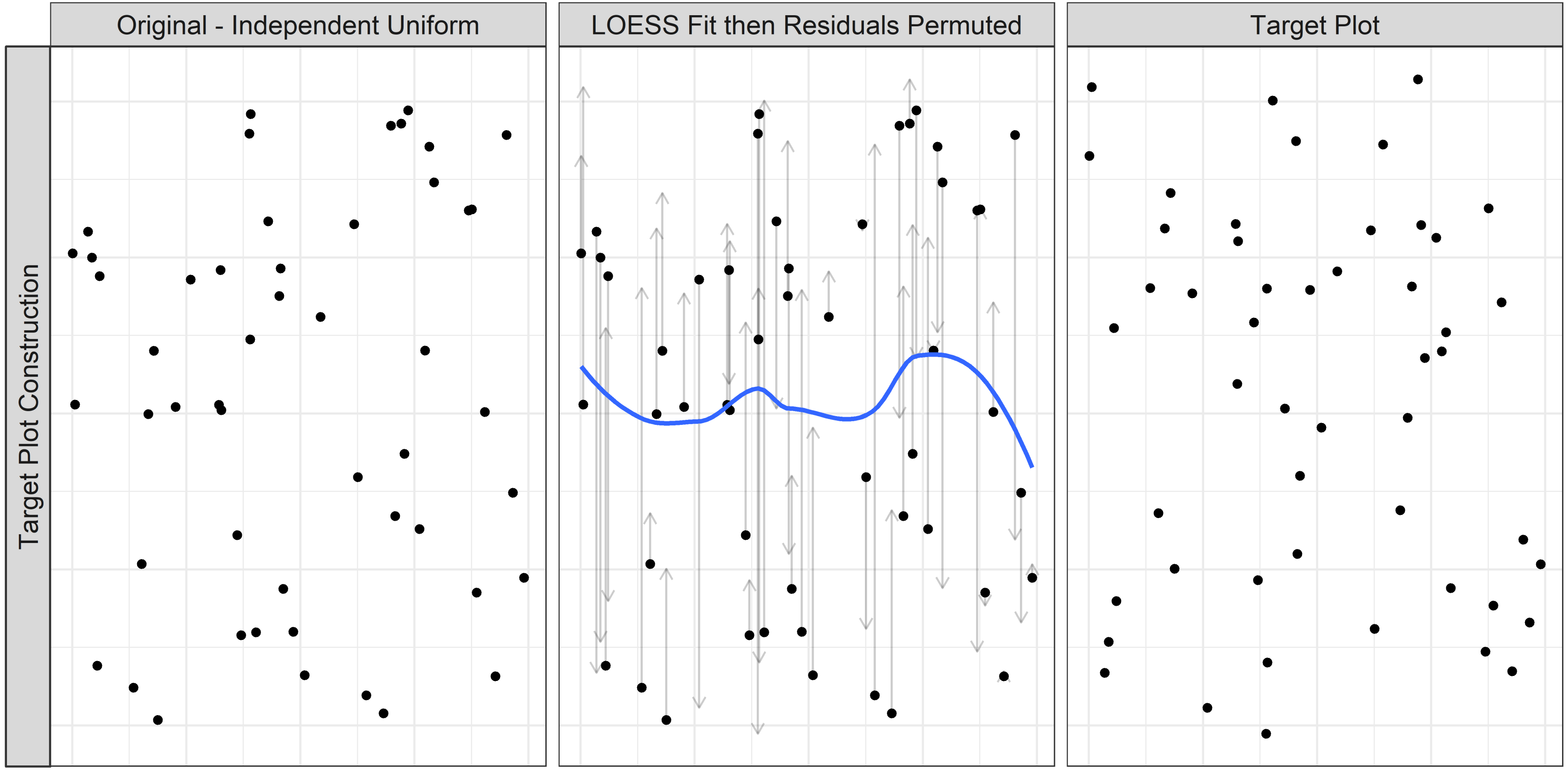
2. Are the TT lineup p-values hypergeometrically distributed? No.

This would be true, **assuming** that null plots and target plots are indistinguishable under the null

So why are target plots noticeably different?

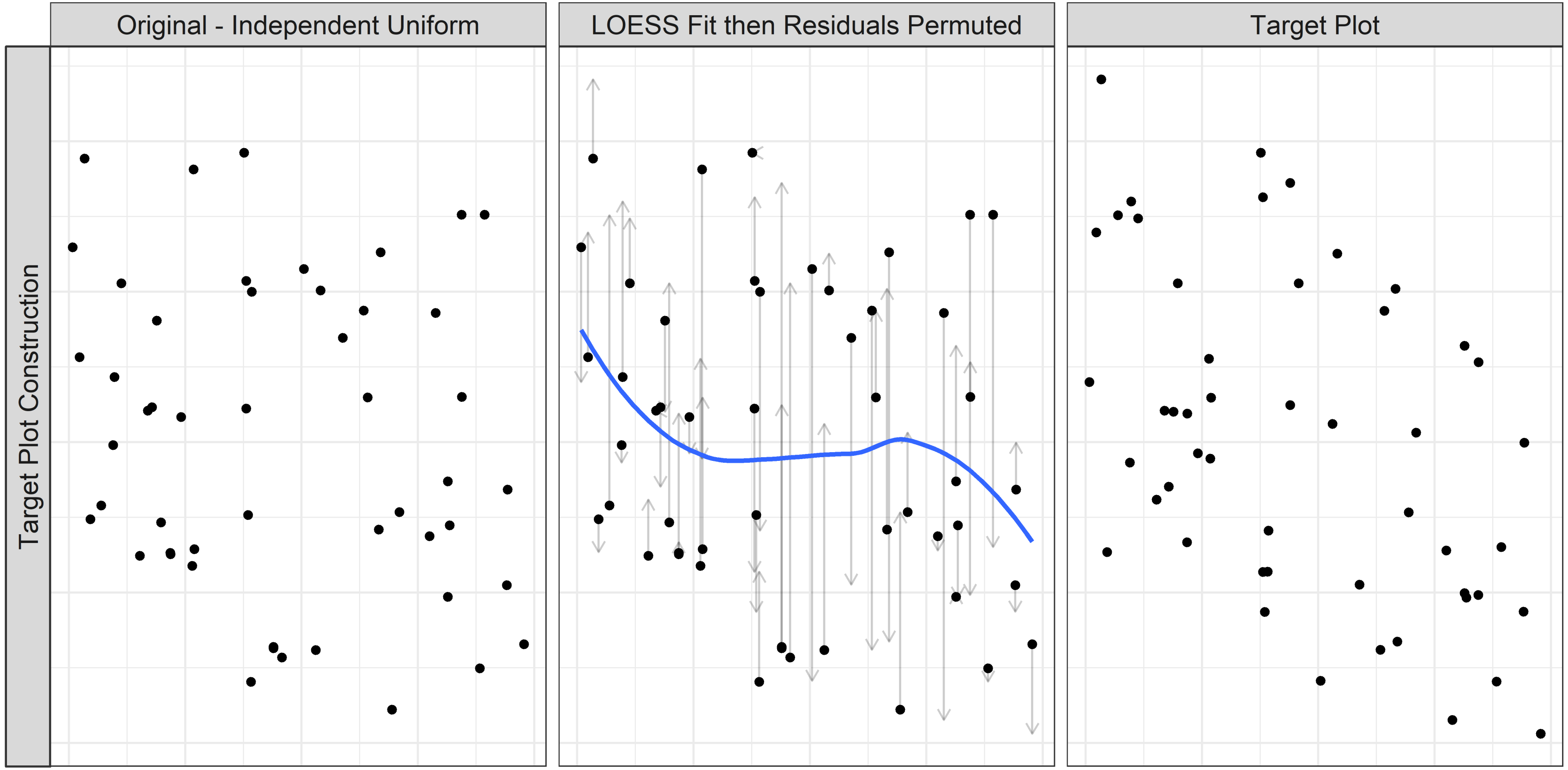
MIAMI UNIVERSITY

2. Are the TT lineup p-values hypergeometrically distributed? No. Why not?



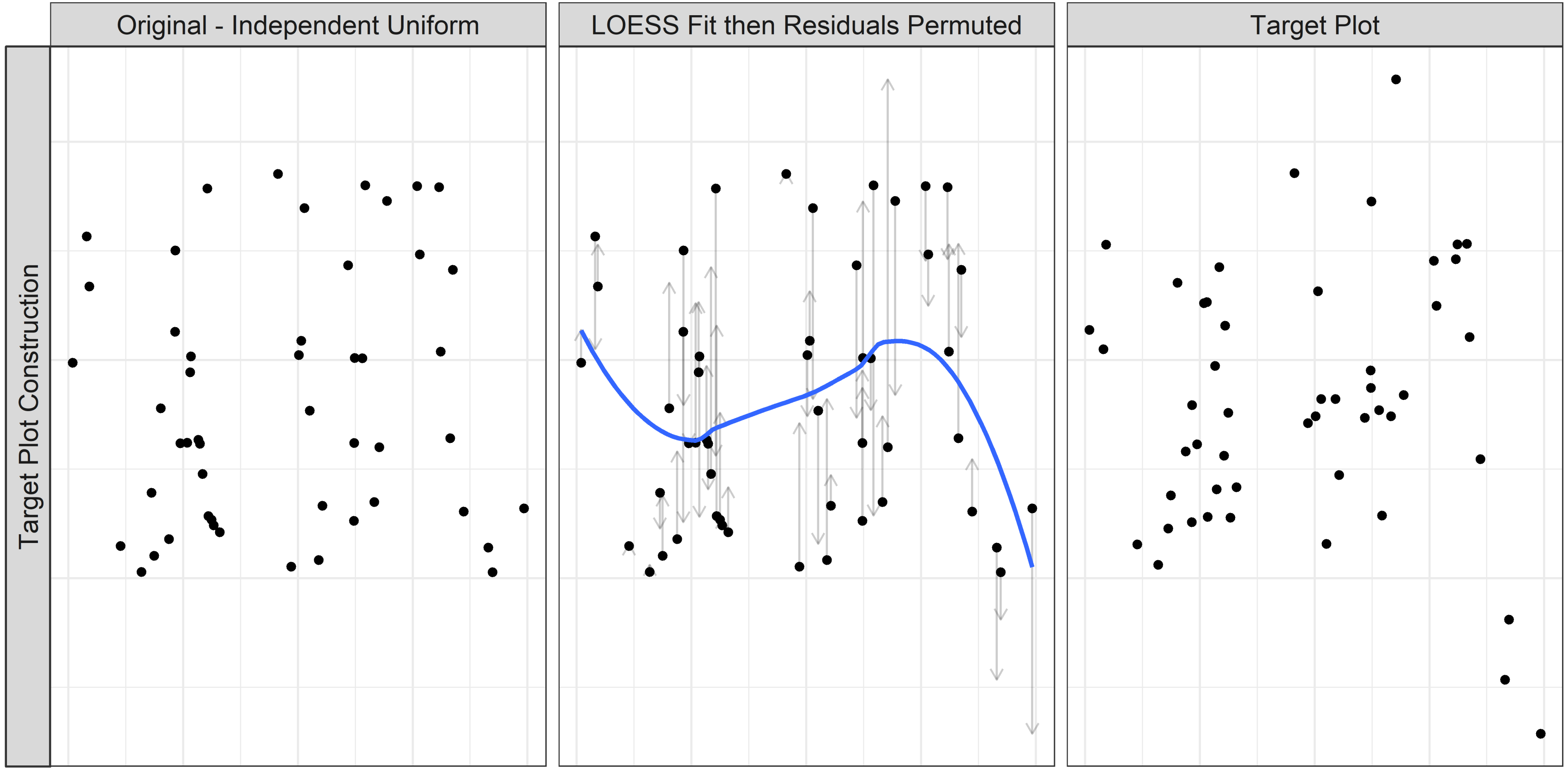
Sometimes data from the null are "steriotypical"

2. Are the TT lineup p-values hypergeometrically distributed? No. Why not?



Sometimes data from the null are "weird" (type I errors)

2. Are the TT lineup p-values hypergeometrically distributed? No. Why not?



Sometimes data from the null are "steriotypical", but our method are unstable

Conclusions

What did we learn?

1. Does the TT lineup have discriminative power in practice?

   Yes, more correct guesses for data with stronger relationship

2. Are correct guess counts hypergeometrically distributed for a true null?

   No, issues with LOESS tail-instability.

   Should consider alternative methods for generating target plots (wider smoothing span or bootstrapping)

Implementation

- R package implementation is available at github.com/kmaurer/teaTasteR

```
devtools::install_github("kmaurer/teaTasteR")
```

MIAMI UNIVERSITY

References

Literature

- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906), 4361-4383.
- Chowdhury, N. R., Cook, D., Hofmann, H., Majumder, M., Lee, E. K., & Toth, A. L. (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. Computational Statistics, 30(2), 293-316.
- Fisher, R. A. (1960). The design of experiments. The design of experiments., (7th Ed).
- Hofmann, H., Follett, L., Majumder, M., & Cook, D. (2012). Graphical tests for power comparison of competing designs. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2441-2448.
- Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. Journal of the American Statistical Association, 108(503), 942-956.
- VanderPlas, S., & Hofmann, H. (2017). Clusters beat Trend!? Testing feature hierarchy in statistical graphics. Journal of Computational and Graphical Statistics, 26(2), 231-242. -Zhao, Y., Cook, D., Hofmann, H., Majumder, M., & Chowdhury, N. R. (2013). Mind Reading: Using an Eye-Tracker to See How People are Looking at Lineups. International Journal of Intelligent Technologies & Applied Statistics, 6(4).

References

Software and Data

- Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39, 357–365. doi: 10.2307/2347385.
- Genz, Bretz, Miwa, Mi, Leisch, Scheipl, Hothorn (2019). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-10. URL http://CRAN.R-project.org/package=mvtnorm
- R Core Team, 2019. R: A language and environment for statistical computing.
- R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.
- Wickham, 2019. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.22.

Thanks!

Slides created via the R package **xaringan**.

MIAMI UNIVERSITY