How sample size impacts Pillai Scores

(and what sociophoneticians should do about it)

Joseph A. Stanley
Brigham Young University
@joey_stan

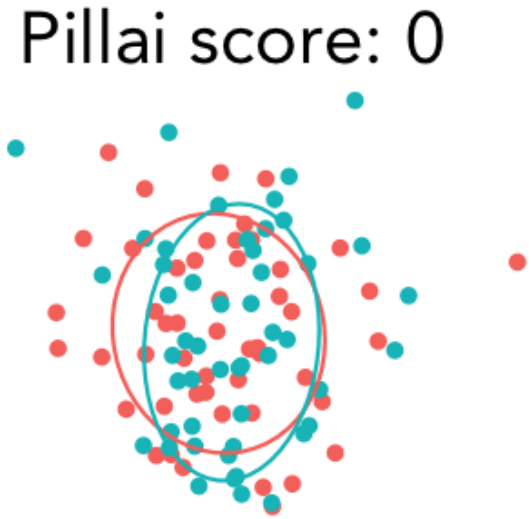
Betsy Sneller
Michigan State University
@betsysneller

New Ways of Analyzing Variation 50
San Jose, California
October 14, 2022

See our paper (under review with in JASA) for more details and code.

Pillai scores in Sociolinguistics

- A way to quantify overlap in multiple dimensions (e.g., F1, F2, duration) (Hay et al., 2006; Nycz & Hall-Lew 2013)
- Some concern about unequal vowel categories sizes (Johnson 2015; convo on Twitter June 2, 2021)
- Our solution: a big simulation to show exactly what impact sample size has on Pillai scores



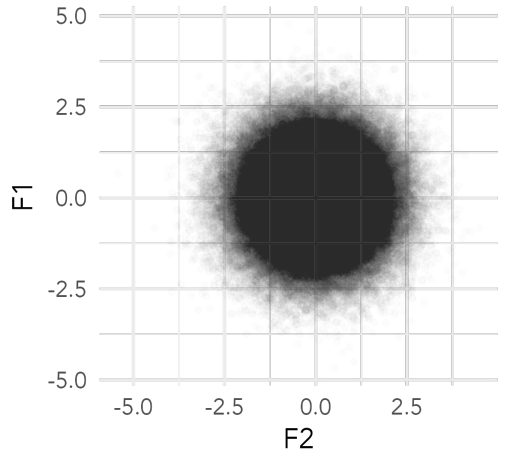
The Simulation

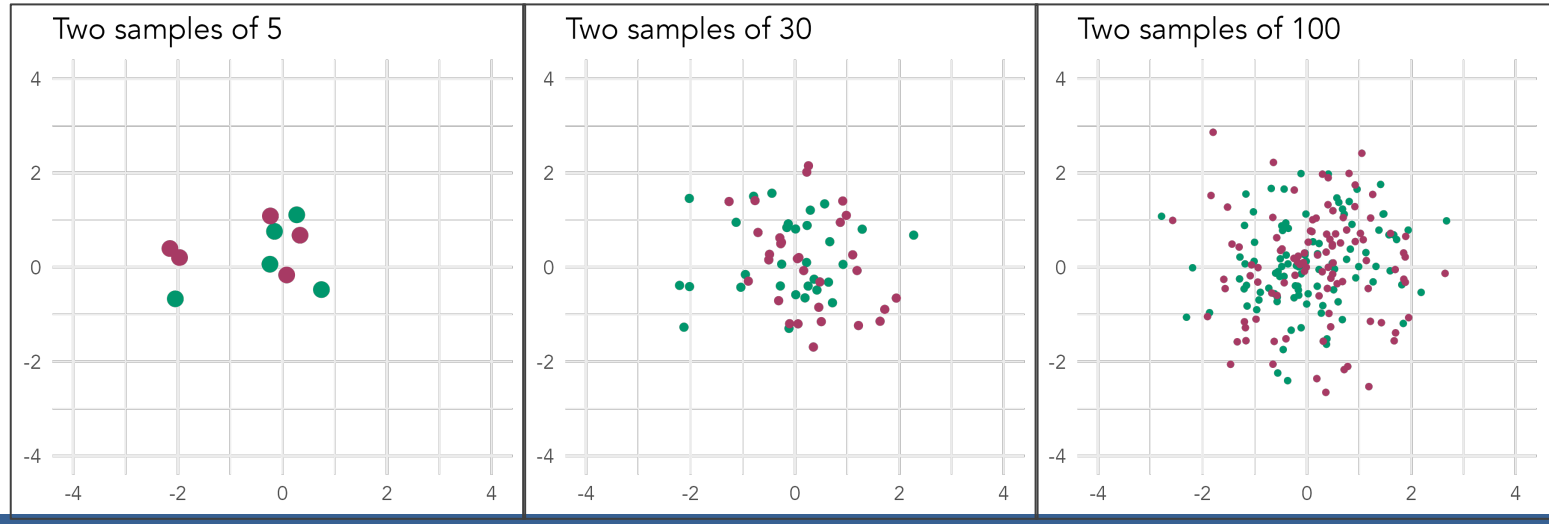
1. Start with a single bivariate normal distribution ▶

2. Sample two "vowel classes" from this distribution. ▼
   - Ground Truth is they're merged.

3. Calculate the Pillai score
   - Should be close to 0 because they're merged!

4. Repeat many, many times

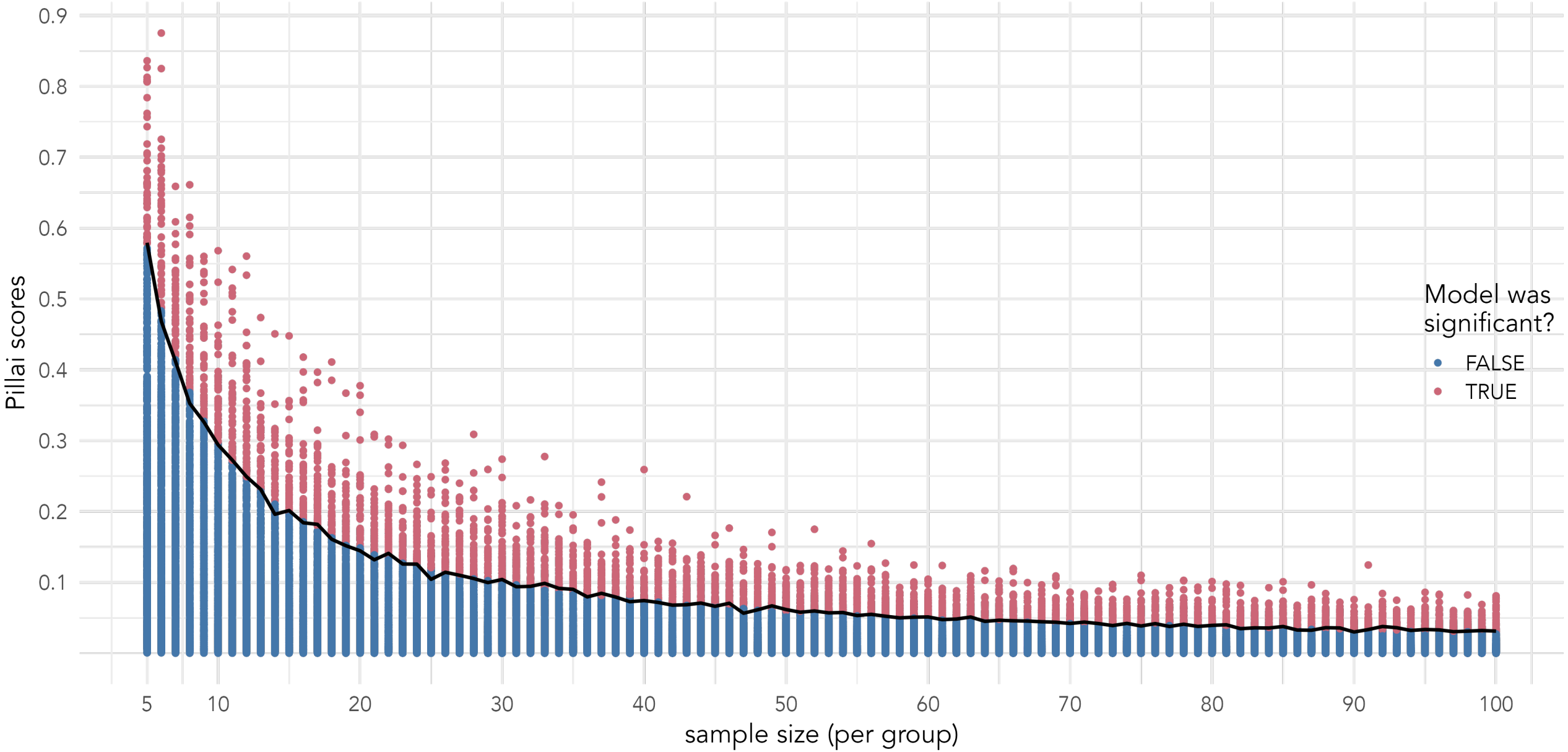




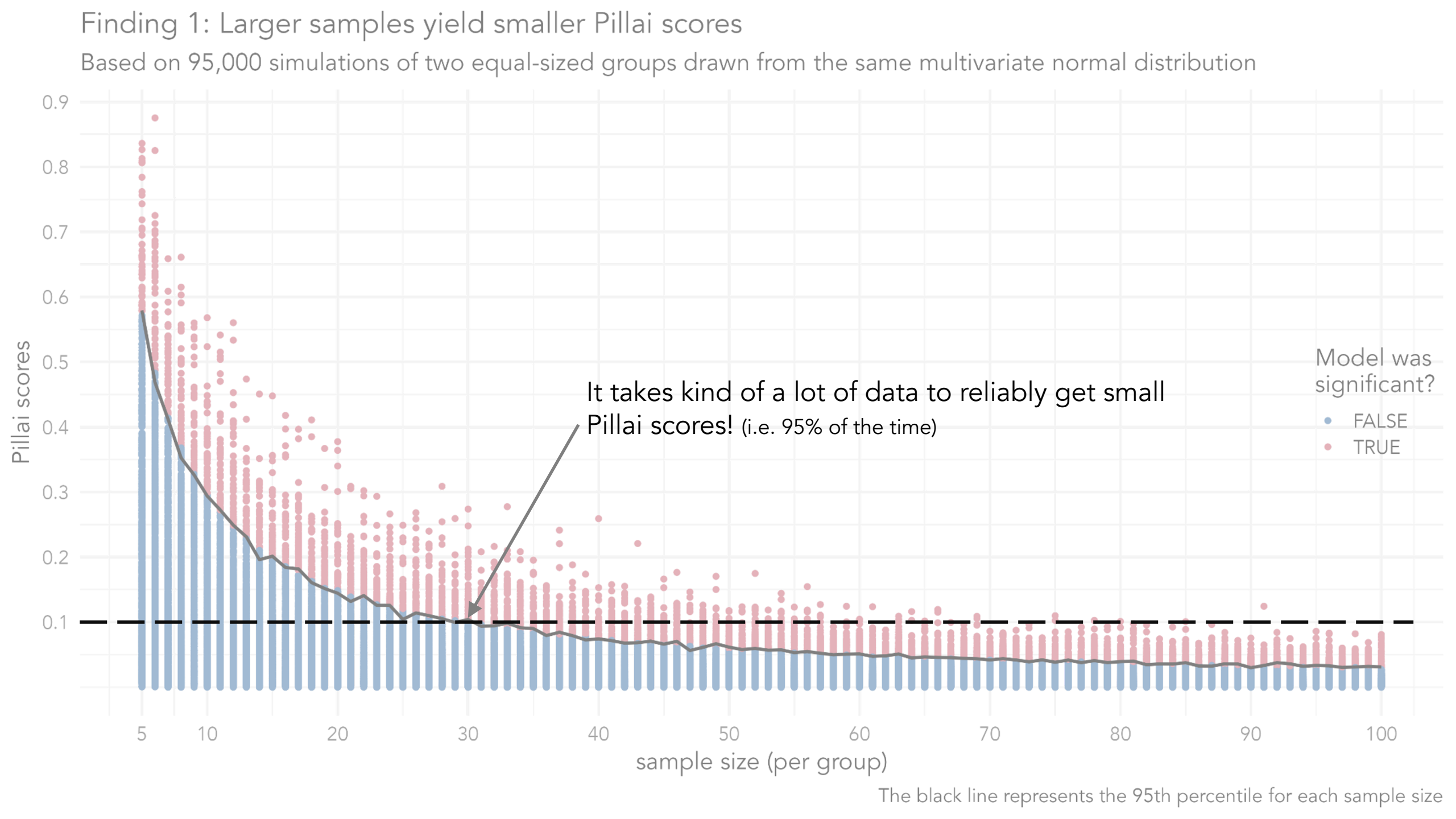
See our paper (under review with in JASA) for more details and code.

Finding 1: Larger samples yield smaller Pillai scores

Based on 95,000 simulations of two equal-sized groups drawn from the same multivariate normal distribution



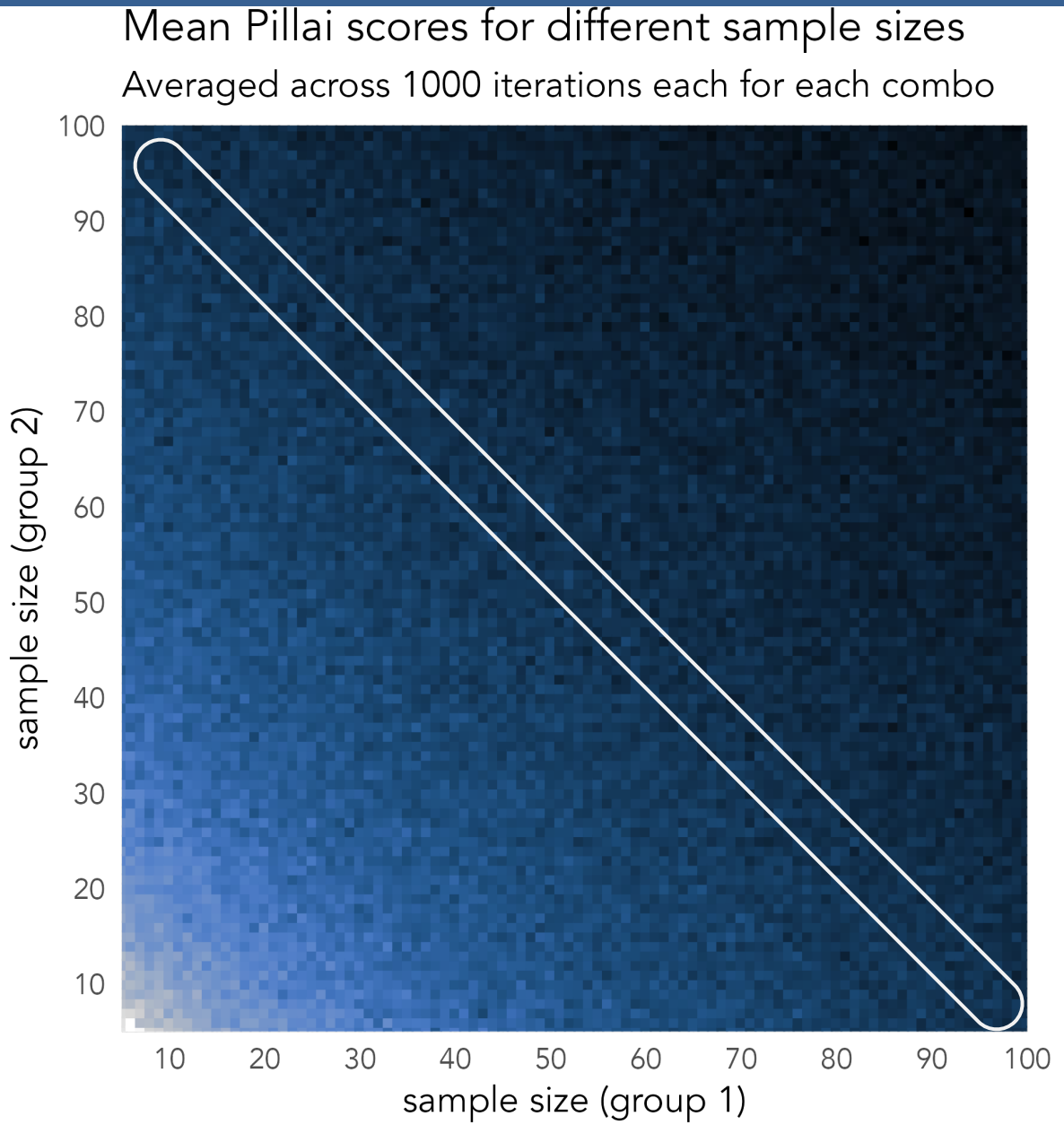
The black line represents the 95th percentile for each sample size


Finding 1: Larger samples yield smaller Pillai scores
Based on 95,000 simulations of two equal-sized groups drawn from the same multivariate normal distribution
Pillai scores
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
It takes kind of a lot of data to reliably get small Pillai scores! (i.e. 95% of the time)
Model was significant?
FALSE
TRUE
5
10
20
30
40
50
60
70
80
90
100
sample size (per group)
The black line represents the 95th percentile for each sample size

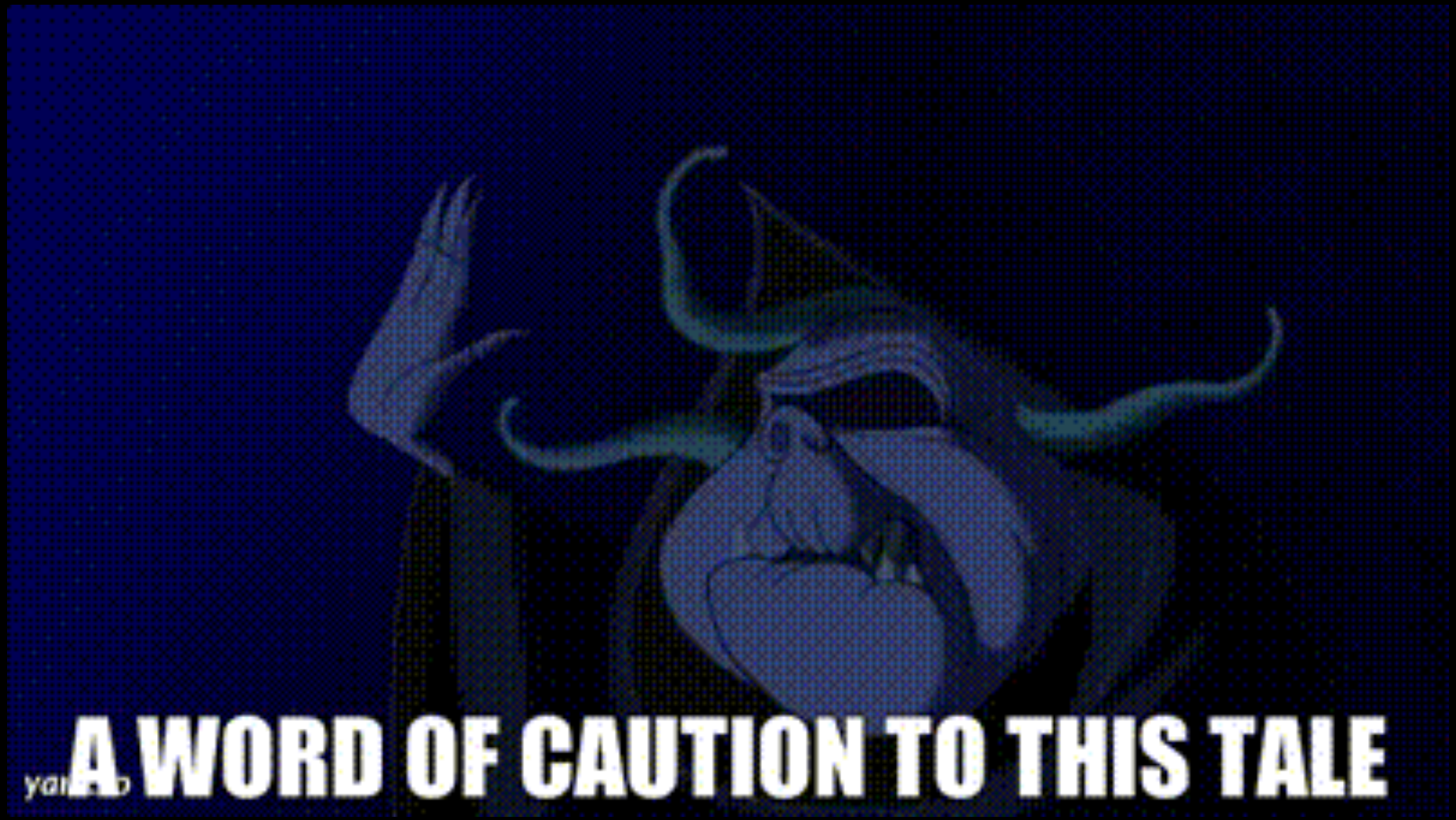
Unequal Groups?

- Sample unequally for each group.
  - Each combo of 5–100 tokens per group, repeat 1000 times.

- Surprising results: unequal sample size doesn't matter! ▶
  - So, what we should consider is total sample size across both vowels
  - Good news for us!



See our paper (under review with in JASA) for more details and code.


A WORD OF CAUTION TO THIS TALE

Warning 1: Don't use the same threshold for all speakers

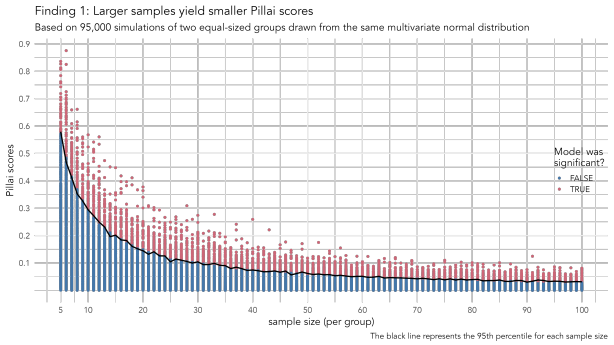
- The threshold for "I'm sure this is merged" should be based on sample size.

- Great news! We've got an equation for you:

$$\text{threshold} = \frac{2e}{n}$$

R code: `2*exp(1)/n`

total sample size



Or you can just use this table ▶

Total sample size	Threshold
10	0.5437
15	0.3624
20	0.2718
25	0.2175
30	0.1812
40	0.1359
50	0.1087
60	0.0906
70	0.0777
80	0.0680
90	0.0604
100	0.0543
200	0.0272
500	0.0109

See our paper (under review with in JASA) for more details and code.

Warning 2: Don't compare to other studies

- Only do so unless you know the sample size of the other study.

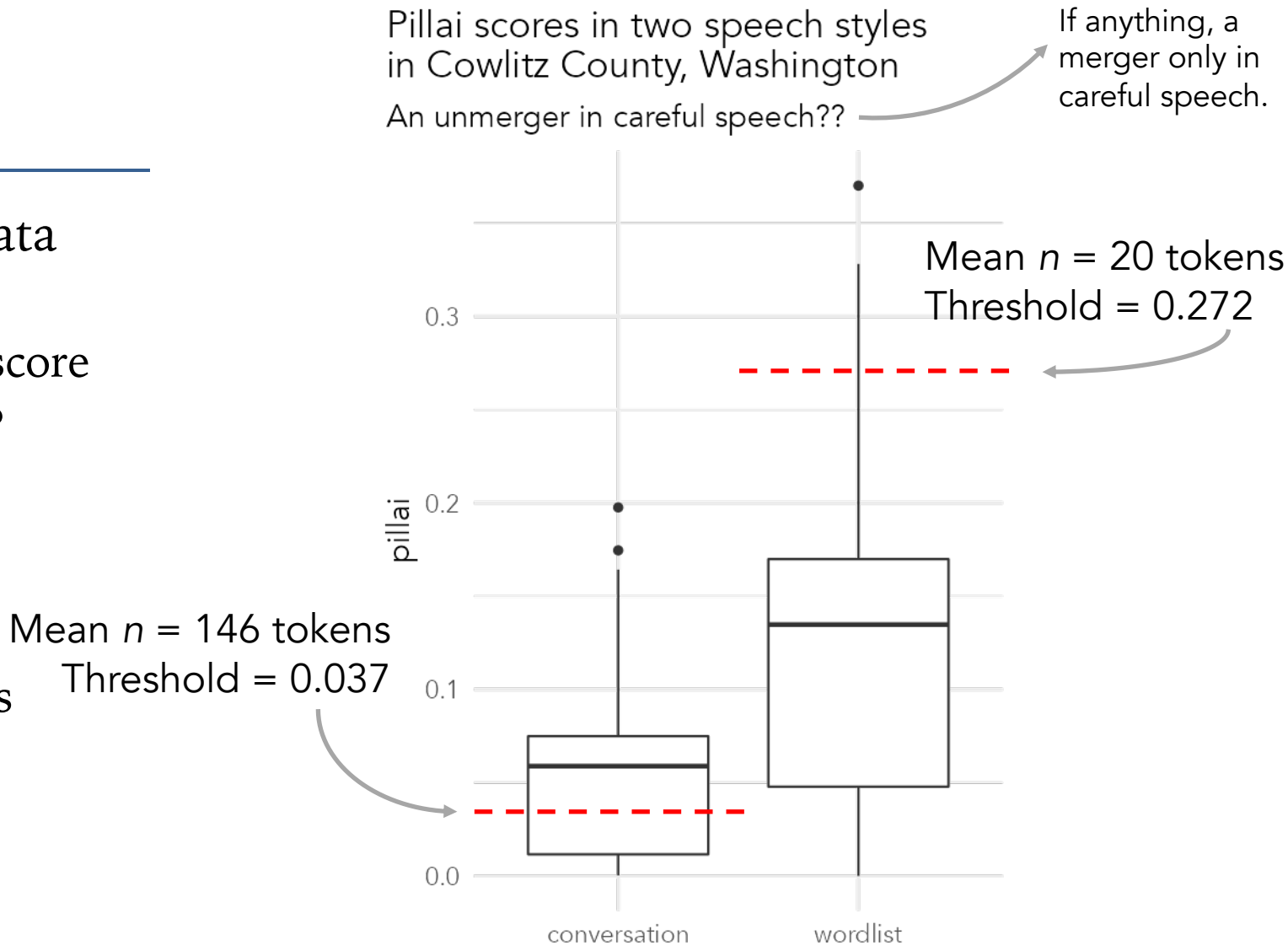
- The solution: report everything to that future people can compare to yours
  - sample size
  - details of the MANOVA model
  - *p*-value from the MANOVA
  - Pillai score
  - Threshold from our formula

"Based on a MANOVA on 179 measurements of F1 and F2, with vowel as the only independent variable, Donna had a Pillai score of 0.0289 (lower than the threshold of 0.0304), with a *p*-value of 0.0756, so we consider her vowels merged."

Warning 3: Careful comparing styles

- Wordlists often contain less data than conversations.
  - Likely to have a higher Pillai score
  - Do we see "unmergers" then?
  - Or is it just math?

- Recommendations
  - Consider (and report) all stats
  - Calculate the threshold
  - Visualize the data



See our paper (under review with in JASA) for more details and code.

Well, then what do we do?

Two basic approaches

- Carefully determine a status ("merged" vs "distinct" vs "unclear"?) for each speaker in each style
  - Pros: a clear understanding of each speaker's data
  - Cons: Discretizing a gradient measure; difficult to analyze change in a speech community over (real and apparent) time

- Normalize for sample size by including *n* in modeling or Monte Carlo simulations
  - Pros: track change over time in a speech community, as distinct categories may become phonetically closer before a true merger takes place
  - Cons: can't tell whether an individual speaker is phonologically merged

- (The approach you use should depend on which question you're interested in.)

See our paper (under review with in JASA) for more details and code.

Final Takeaways

- Sample size matters! But only total sample size (across both categories)
  - Report sample size! And p-values! And everything else!
  - Use our threshold suggestion if trying to determine “merged” vs “distinct” for an individual speaker
  - Or our normalization suggestion if trying to track a change towards or away from a merger over real/apparent time

How sample size impacts Pillai Scores

(and what sociophoneticians should do about it)

Joseph A. Stanley
Brigham Young University
@joey_stan

Betsy Sneller
Michigan State University
@betsysneller

Template for reporting Pillai scores ▼

"Based on a MANOVA on 179 measurements of F1 and F2, with vowel as the only independent variable, Donna had a Pillai score of 0.0289 (lower than the threshold of 0.0304), with a p-value of 0.0756, so we consider her vowels merged."

Threshold formula ▶ $\frac{2e}{n}$

R code ▶

```
2*exp(1)/n
```

Thresholds at various sample sizes ▶

Total sample size	Threshold
10	0.5437
15	0.3624
20	0.2718
25	0.2175
30	0.1812
40	0.1359
50	0.1087
60	0.0906
70	0.0777
80	0.0680
90	0.0604
100	0.0543
200	0.0272
500	0.0109

See our paper (under review with in JASA) for more details and code.