Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

Lecture 1: Statistical problems

Slides adapted from Andy Wedel, Bodo Winter

About me

The LAnguages, Machines & Brains Lab (LAMB) is directed by Dr. Jixing Li. Our lab's research intersects the multidisciplinary fields of Linguistics, Natural Language Processing (NLP) and Cognitive Neuroscience. We are always looking for motivated graduate and undergraduate students, postdocs, research assistants and visiting scholars to join our team. If interested, please email jixingli@cityu.edu.hk.

Office: LI-5459
Schedule office hours with me if you need help.

lamb

LAnguages Machines and Brains

Home | People | Research | Publications

Welcome to the LAnguages, Machines & Brains Lab (LAMB)

Lecture plan

- The course logistics
- Three statistical problems
- R and R Studio
- Short break (15 mins)
- Hands-on exercises

Course logistics

- Instructor: Jixing Li
- Location: LI-4307
- Time: Tues 3:00-5:50 pm HKT

- Check Canvas for announcements, labs, assignments, slides, recordings etc. Slides and recordings will be uploaded after each lecture.

- Textbook (not required reading): Winter, Bodo. *Statistics for Linguists: An Introduction Using R*. New York: Routledge, 2020

Course work and grading policy

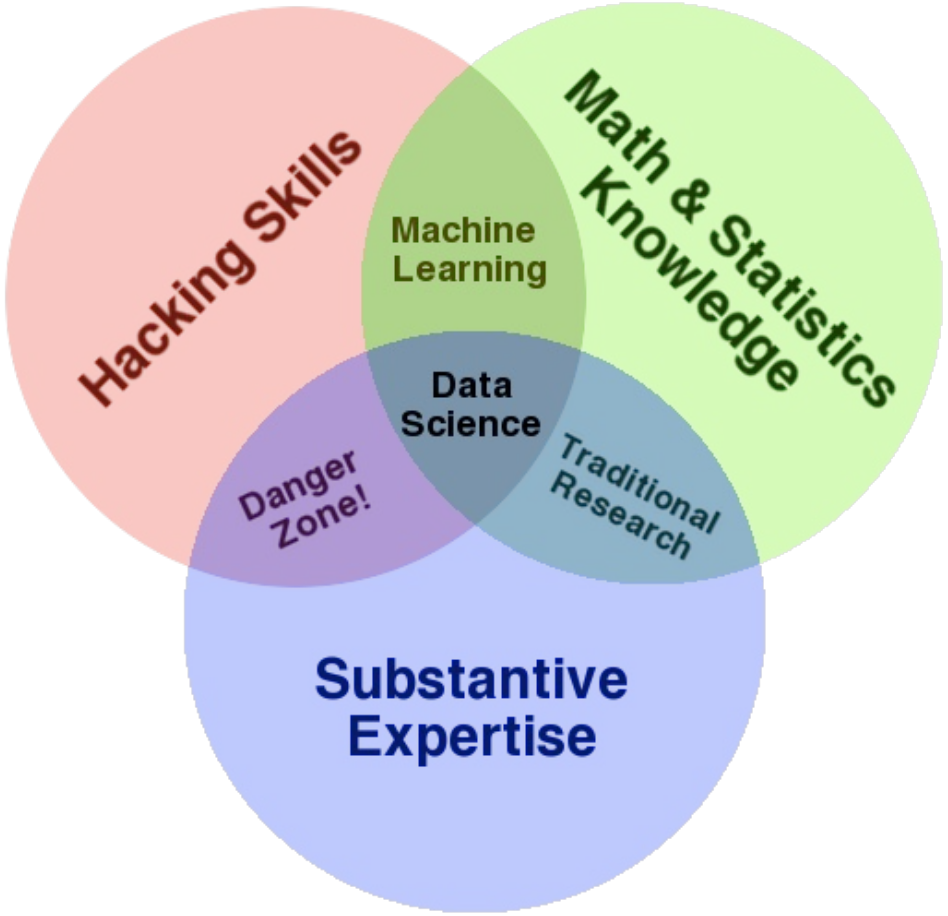
- 3 assignments: 3 x 15%
  - Submitted to Canvas
- Oral presentation (group of 2): 15%
  - Last class
- Final exam: 40%
- Attendance: 3 free absences; afterwards, 1% off the overall course grade.

“Data is merely the raw material of knowledge”

“The big problem is going to be the ability of humans to use, analyze and make sense of the data.”

http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1&

The "data science" rebranding



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

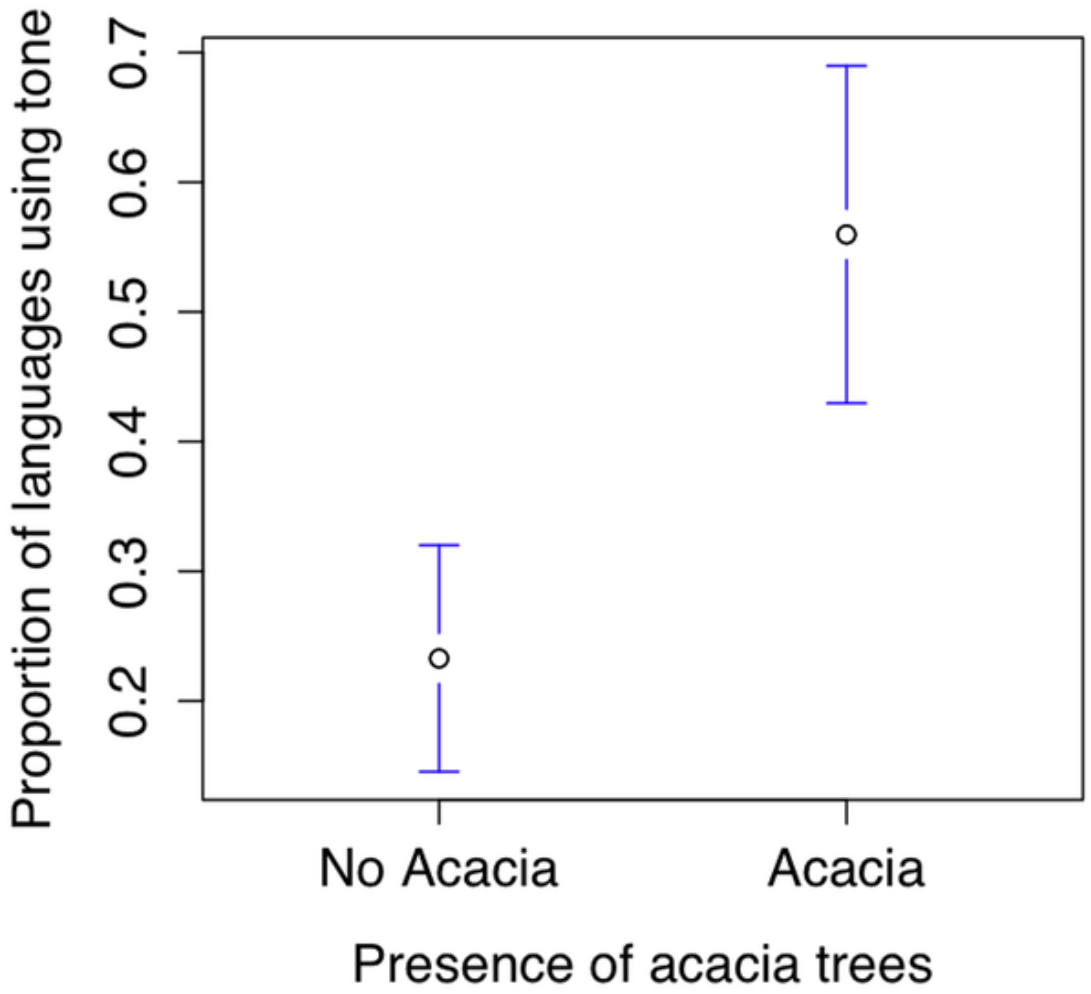
Three statistical problems

Correlation is not causation

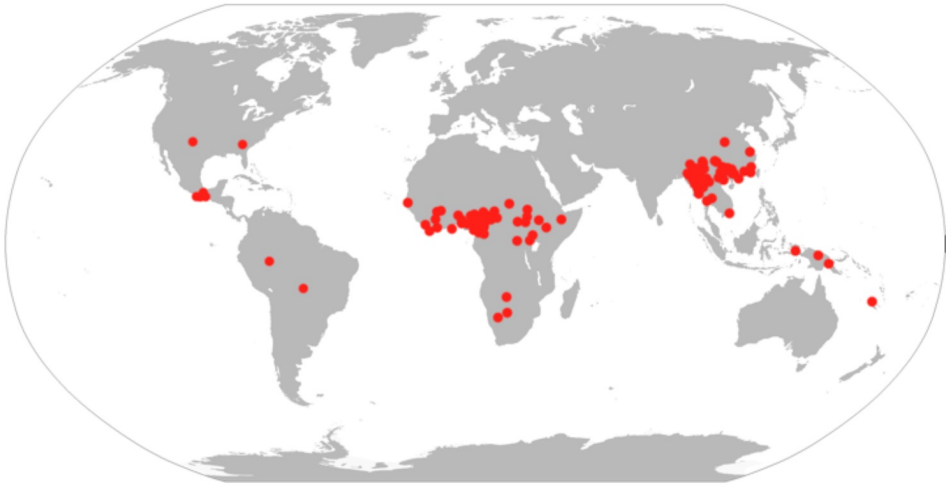
Multiple comparisons

Lack of independence

Acacia trees and tonal languages

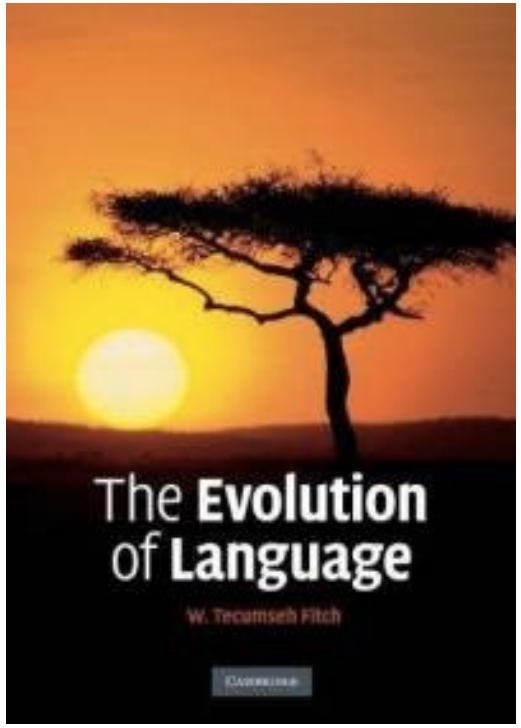


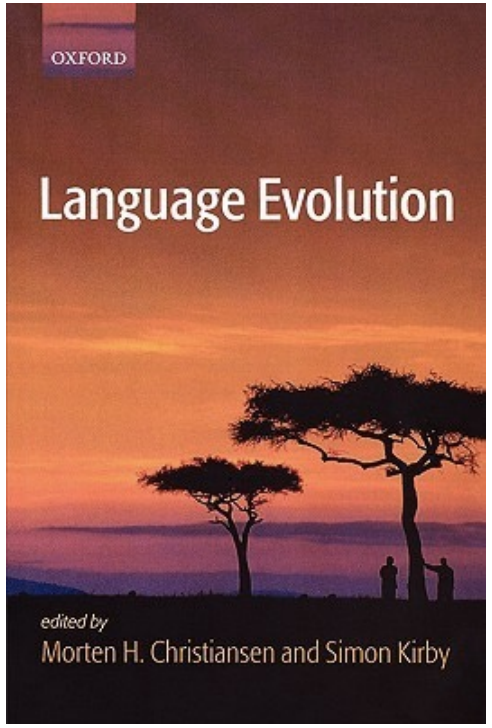
Distribution of acacia trees


Distribution of tonal languages

Roberts, S., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS one*, 8(8), e70902.

Acacia trees and tonal languages

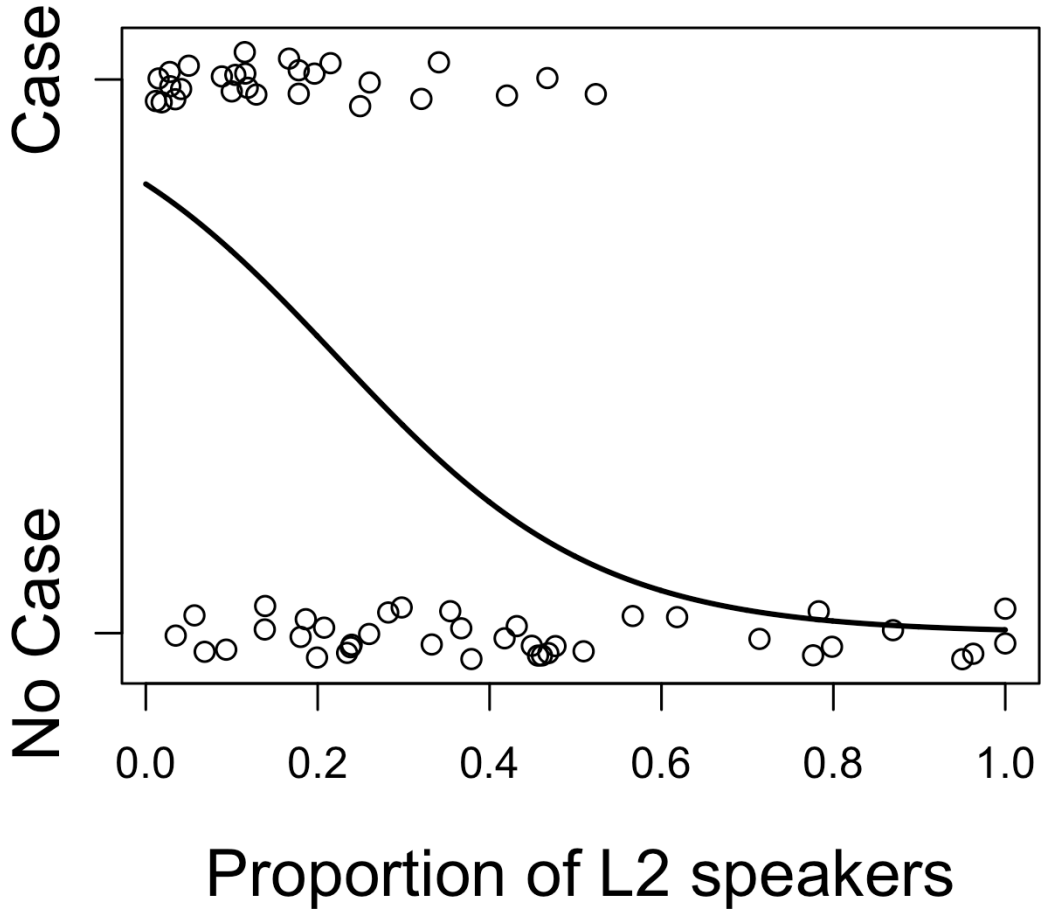




Acacia trees
→ Dimethyltryptamine (DMT)
→ Some primates got high
→ Language evolution

https://specgram.com/CLXII.4/07.geraint.acacia.html

L2 learners and Case in language



Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics & Change*, 3:1, 1-27.

1. L2 learning → morphology

2. L2 speakers ← morphology

3. L2 speakers ↔ confound? ↔ morphology

Steps to support causality

The data must be strong.

The data must be consistent.

The data must be coherent.

The data must be specific.

The causal effect must be plausible.

http://www.skepticalraptor.com/skepticalraptorblog.php/correlation-does-not-imply-causation-except-when-it-does/

Three statistical problems

Correlation is not causation

Multiple comparisons

Lack of independence

“The dead salmon” study

A cluster of activated voxels have been identified in the brain of a dead salmon that had been “asked” to perform a social perspective-taking task while lying inside a fMRI scanner.

https://www.sciencenews.org/article/trawling-brain

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, 1, 1-5.

Example of doing 100 tests:

If $\alpha = 0.05$ is taken as the significance level

For 100 tests, the expected number of incorrect rejections of the null hypothesis is 5

The probability of *at least one* statistical result being significant is:

$$1 - 0.95^{100} = 0.994$$

What to do?

(1) Correcting for multiple comparisons (e.g., Bonferroni correction)

(2) Avoid multiplicity at the design stage (Bender & Lange, 2001)

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how?. *Journal of clinical epidemiology*, 54(4), 343-349.

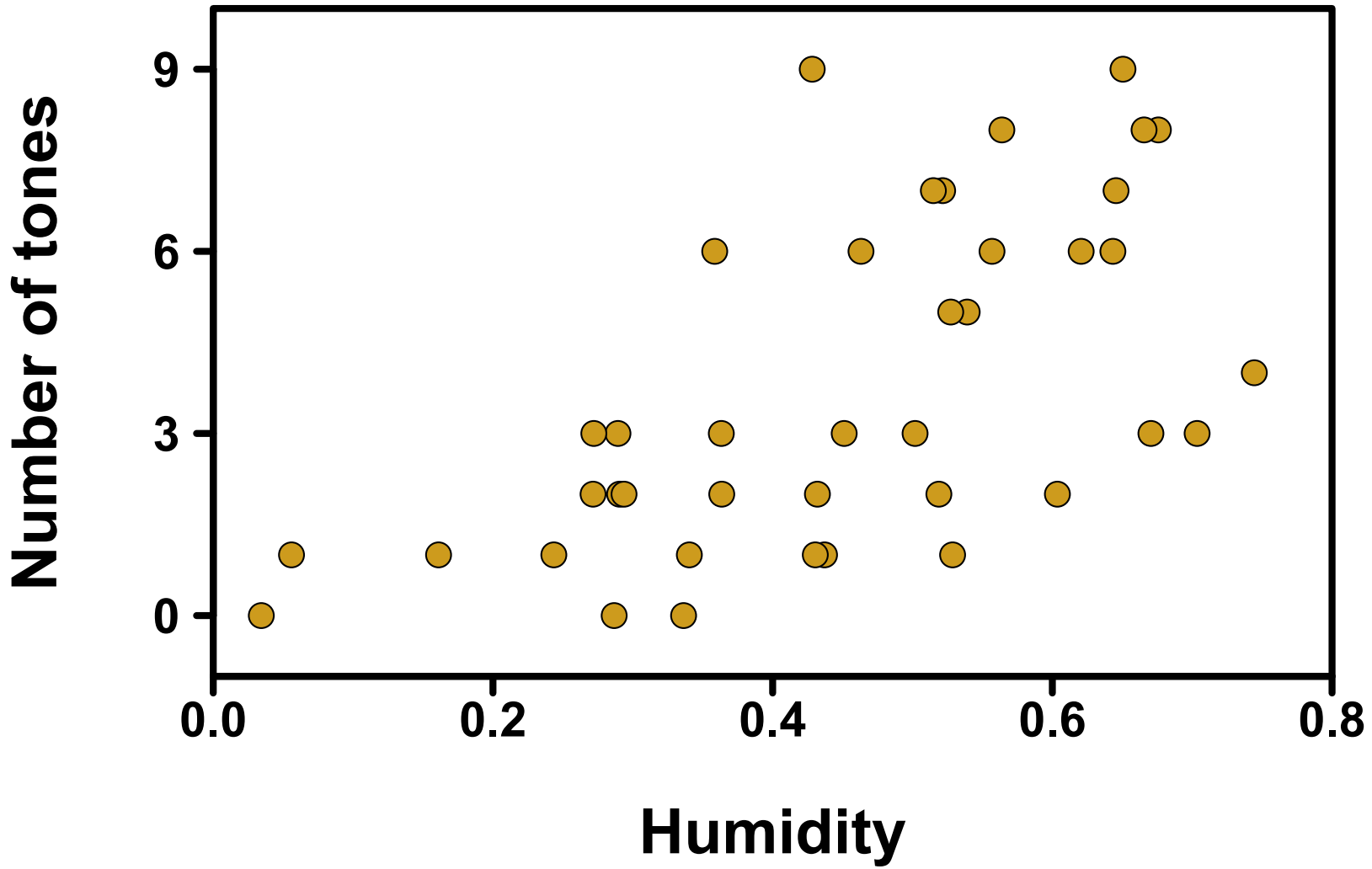
Three statistical problems

Correlation is not causation

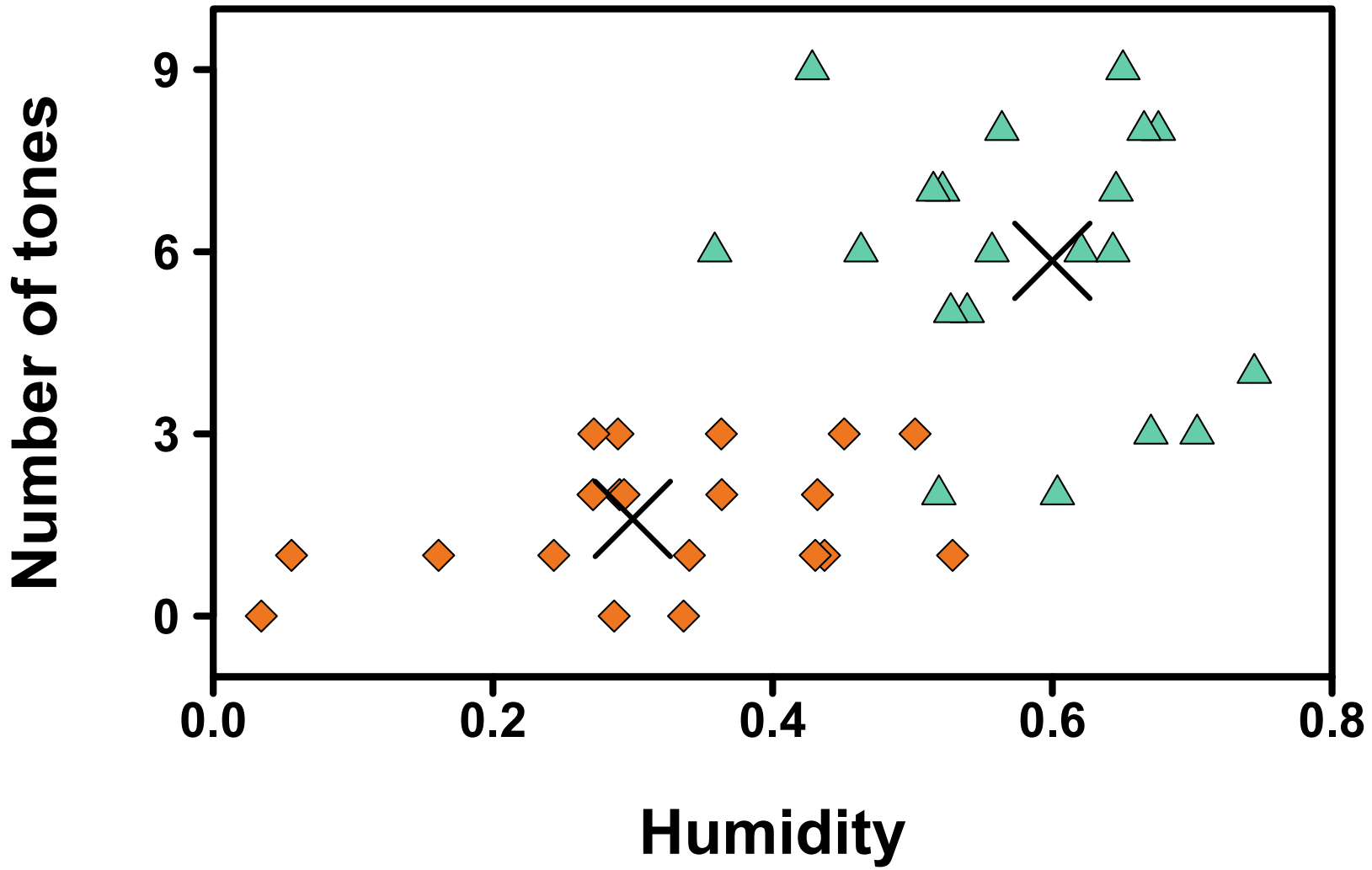
Multiple comparisons

Lack of independence

Humidity and number of tones



Humidity and number of tones



Contact relationships

"If you borrow your writing system from the Chinese, it's likely that you'll borrow some aspects of food culture and architecture as well. This is not because a logographic writing system tends to cause people to use chopsticks or build pagodas."

Mark Liberman

http://languagelog.ldc.upenn.edu/nll/?p=3764

To do

- Install R and R Studio on your laptop
- Finish Lab 1