Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington
Week 9: 5/21-5/23

Overview

- Conceptual discussion on Monday (today)
- Statistics with a focus on exploratory techniques
- Parametric vs. non-parametric tests
- Working with multiple datasets
- Practical work on Wednesday Lab 8 (on practicum group data):
    - come with textfiles of your data

Reading questions

- Are there any papers that use mixed effects models in sociolinguistics?
- Are there any cases in linguistics where we would use something other than individual as a random effect?
- Linguistics articles exemplifying "Random Forests"?
- Interactions between variables: are there good ways of quickly teasing out the relevant variables and chains of variable dependence?
- t-tests and chi-square tests: Suitability or overuse?
- Varbrul: Are there any benefits to using multiple approaches first and then running a Varbrul analysis, rather than going straight to another package.

Guilty until proven...

- Non-parametric tests are often feared. Why?
- We should probably use them more in investigations of our own original research.
- We should almost always use them in studies with secondary data or multiple corpora.

Descriptive vs. Inferential

- Descriptive statistics -- allow us to **understand** the properties of our data (exploratory statistics).
- Inferential statistics -- used when we have a research question, formulated hypotheses about our data. Tell us whether the alternative hypothesis about our data is likely to be closer to "truth", help us to **interpret** our results relative to our RQ.

Assumptions of Parametric Tests

- Normally distributed data--symmetrical about a mean
- Homogeneity of variance--variances are same throughout the data
- Interval data should be measured at the interval level (equal distances on the scale are given equal differences in values)
- Independence--one subject's data is not influenced by another

Large Database Considerations

- You didn’t collect the data (even token count of variable of interest may be unknown)
- You don’t know which assumptions are satisfied and which aren’t
- Your corpora are known to be different in size, central tendency, variance, sampled from different populations* using different measurement techniques...

...probably need non-parametric methods,

...definitely need exploratory work first

Terms

- Population: In statistics, the set of all potential entities or values, including not only cases actually observed but those that are potentially observable.
- Sample: A few units drawn from the population
- Dependent variable(s): The variable(s) you are trying to understand
- Independent variable(s): The factor(s) that might potentially affect the outcome.

Variable types

Both Independent and Dependent variables may be:

- categorical - entities may be grouped into discrete, named classes or categories. If these categories are named, we have a nominal variable. These are often dichotomous/binary:

  - e.g., "male, female"

- ordinal - entities that can be grouped into numbered categories, but with more information value than the nominal variables, because the numbers may be taken to indicate rankings, or ordered relations:

  - e.g., SyntComplexity (pron NPs "1" < simple lexical NPs "2" < non-clausally modified lexical NPs "3", clausally-modified NPs "4")

- continuous - value lies on some scale

  - e.g., duration values, formant values, pitch, etc.

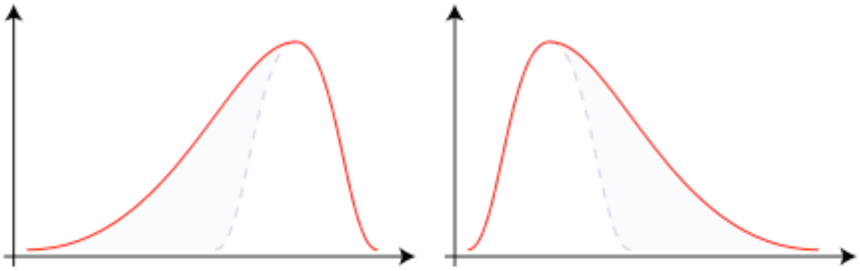
examples

- Stress (stress vs. unstressed)
- StressPosition (1st syllable, elsewhere)
- Idiomaticity of a VP (1 = high or completely idiomatic, 2 = intermediate or metaphorical, 3 = low or fully compositional)
- NPType (lexical vs. pronominal)
- Animacy (animate vs. inanimate)
- Euclidean distance
- Burst (released vs. unreleased)

	Dependent	Independent
Histogram	single, continuous or categorical	n/a (relative freq of some value)
Boxplot	single, continuous	categorical
Scatterplot	continuous	continuous

Normality

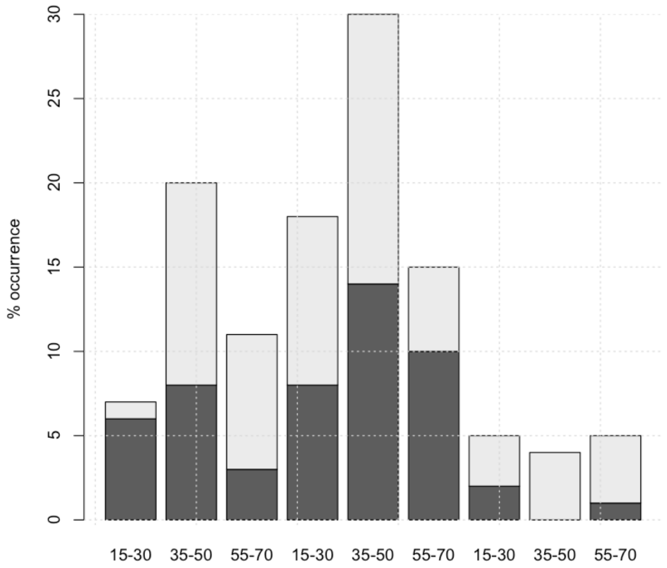
- Frequency table:
  - range: minimum and maximum, range
  - central tendency: median, mean, mode
  - variability: standard deviation, variance, quartile range
  - shape: skewness and kurtosis



Negative Skew

Positive Skew

```
    token          store          style            word          r_present
Min.   :  1.0    Klein:216    casual   :456    4th  :380      0:499
1st Qu.:182.5    Macys:333    emphatic:271    floor:347      1:228
Median :364.0    Saks :178
Mean   :364.0
3rd Qu.:545.5
Max.   :727.0
```

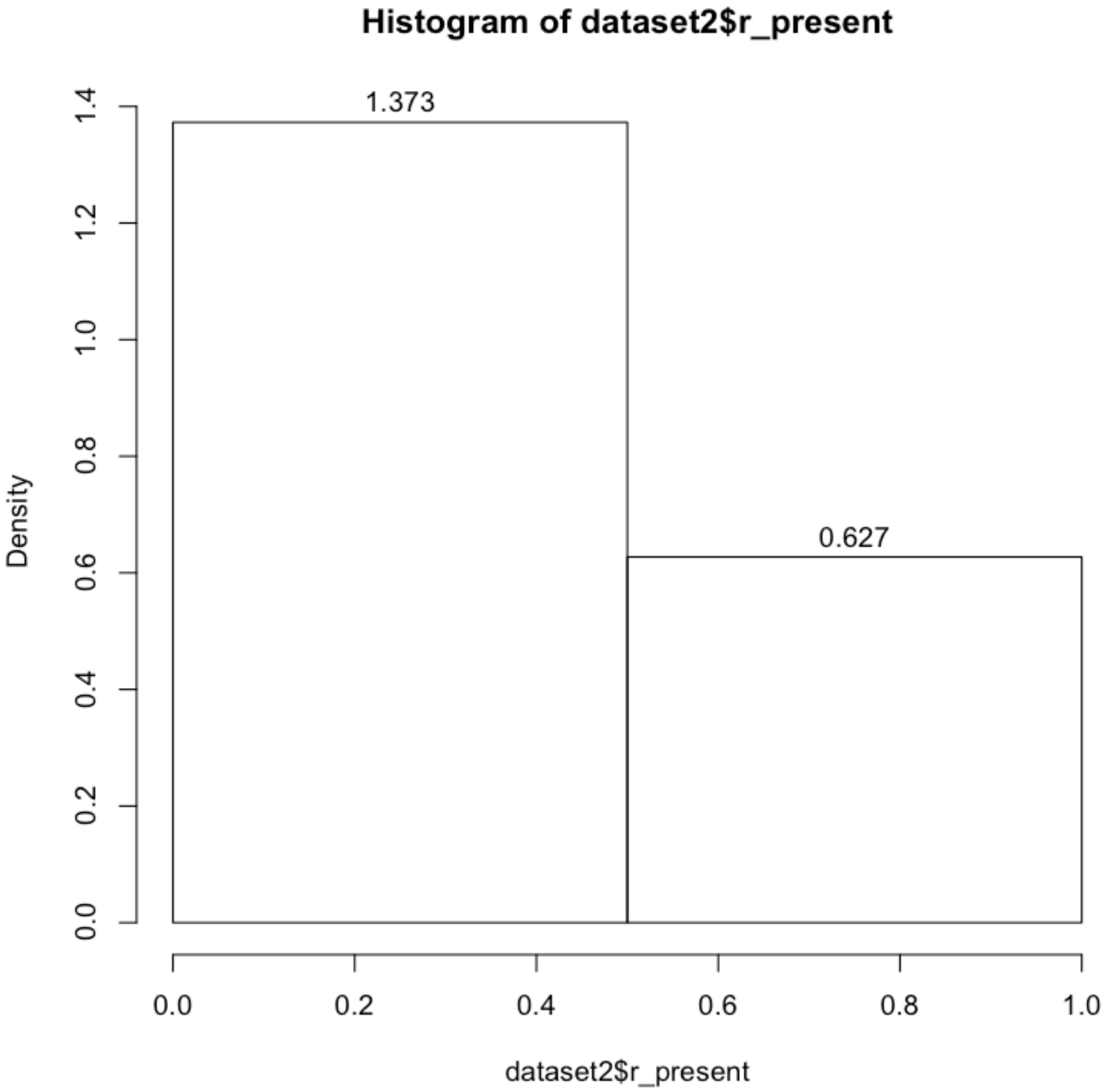


Normality

- Histogram (aka “frequency distribution”. Types: simple, stacked, frequency polygon)
  - shows frequency with which the values of the dependent variable occur
  - should add up to 100%
  - doesn't tell us the mean
  - different from bar graphs in that bar graphs show just the means (not all the observed values) for your grouping variables

Normality

- Histogram
  - Cumulative frequency plots (P-P or probability-probability plots) sort the probability of our variable against the cumulative probability in a normal distribution, and calculate ranks for these using z-scores. Can tell us how likely we'd be to see a value if the distribution were normal.


Histogram of dataset2$r_present
1.373
0.627
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
dataset2$r_present

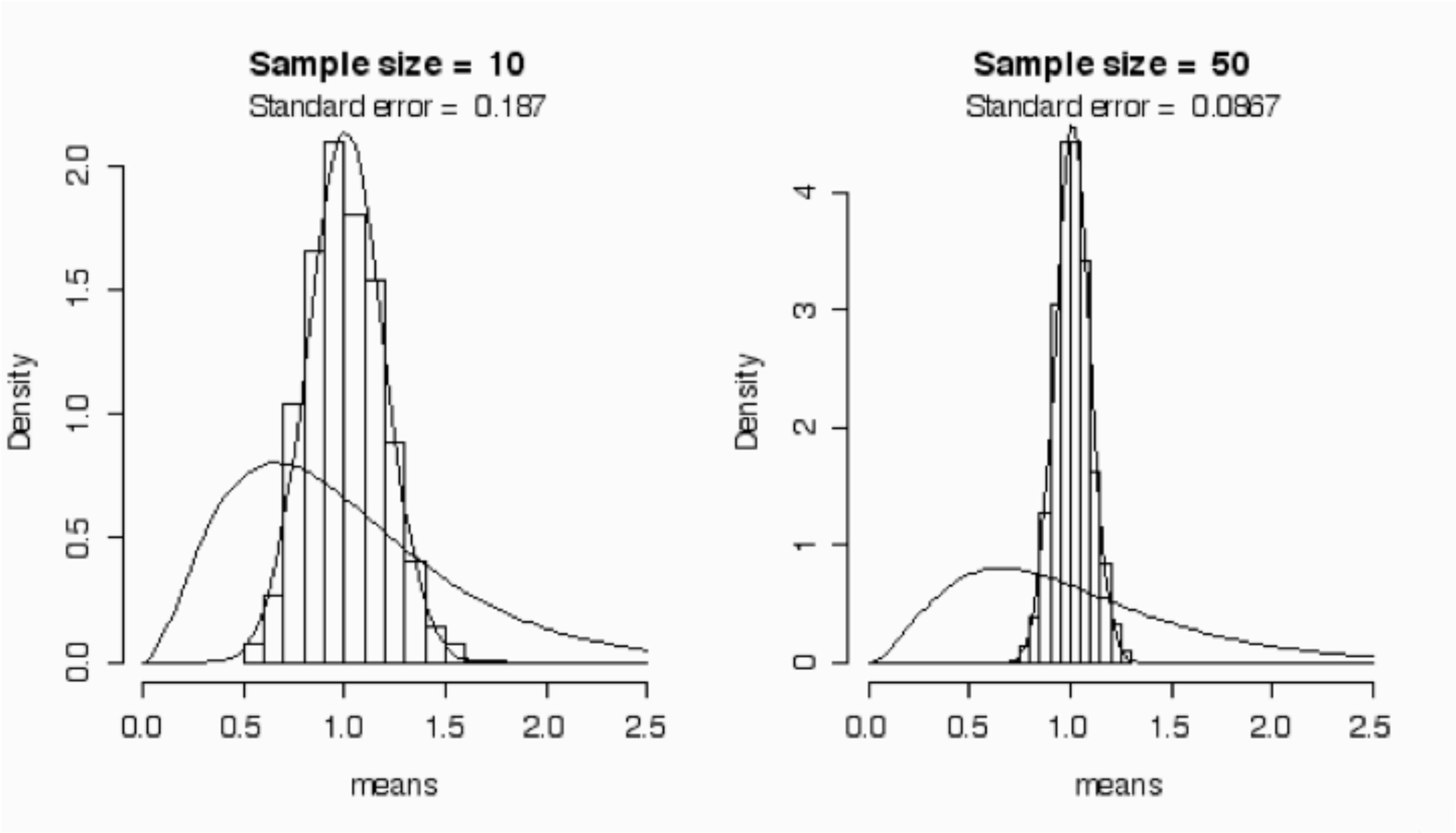
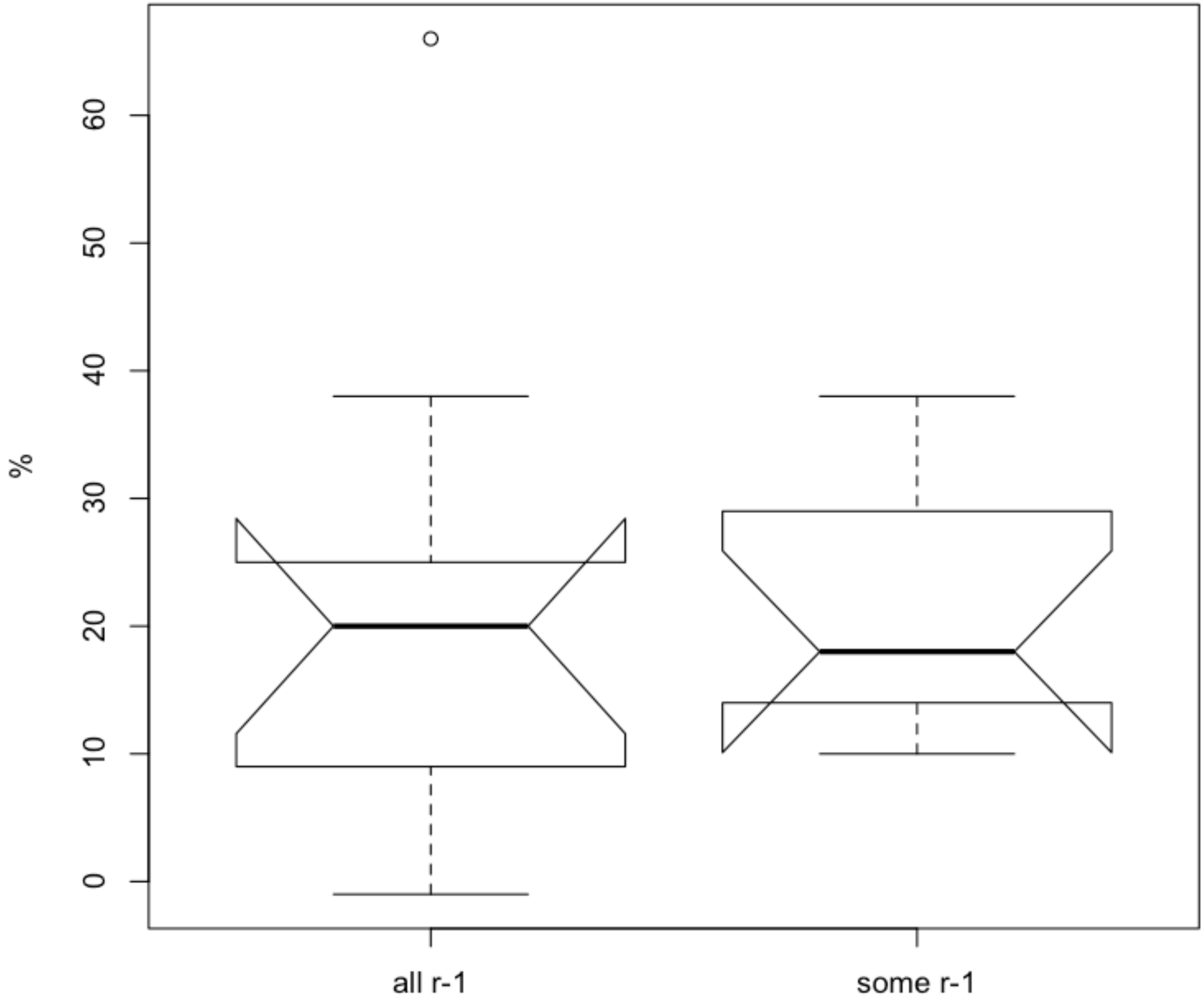


Figure 2.3. The sampling distribution of the mean taken from 1000 samples that were drawn from a skewed population distribution when the sample size was only 10 observations, and when the sample size was 50 observations.

Johnson (2008)

Normality

- Box plots (simple, clustered)
- Show the range of the scores or values in the data
- Tell us whether the distribution is skewed or symmetrical about the median
- Median displayed at center of box marking off the interquartile range
- Interquartile range shows the middle 50% of observations
- Optional whiskers extend to the highest and lowest values in the data


%
0
10
20
30
40
50
60
all r-1
some r-1

Scatterplot

- absolutely should come before reporting a correlation test
- can highlight the parametric or non-parametric distribution of the data
- use for plotting continuous dependent variables against continuous independent variables

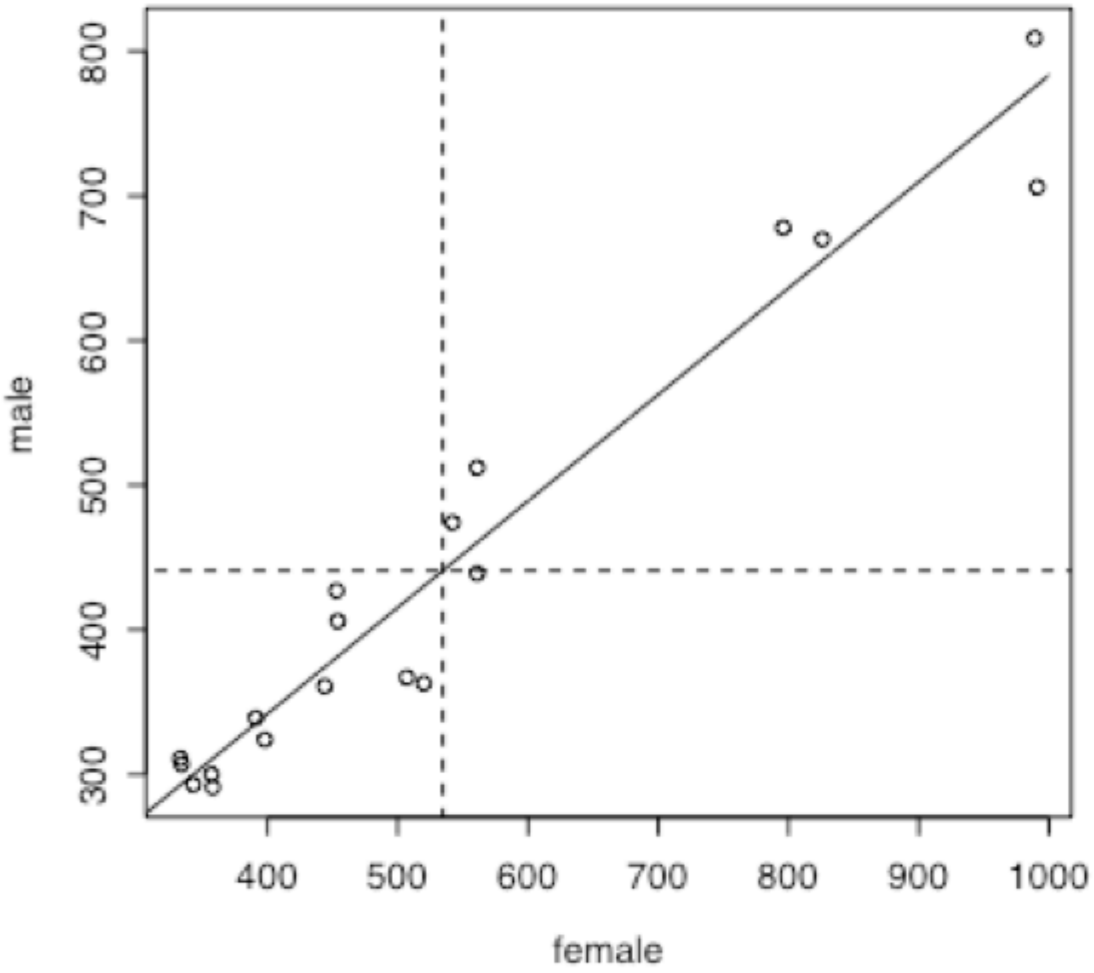


Figure 2.7. Nineteen pairs of male and female F1 values drawn from four different languages and 4 or 5 vowels in each language. The grid lines mark the average female (vertical line) and male (horizontal line) F1 values. The diagonal line is the best fitting straight line (the linear regression) that relates female F1 to male F1.

Johnson (2008)

Multiple Sets of Data

- You can usefully examine these properties in all the datasets you choose to use.
- Don't combine the datasets unless you can specifically show they're drawn from the same population.
- Kolmogorov-Smirnov and Shapiro-Wilk: specifically test whether sample scores come from a normal distribution (*p*<.05 here indicates a *non-normal* distribution)

Homogeneity of variance

- as you explore the structure of all levels of your variable, the variances are the same.
- look at variances and check, or
- use Levene’s test (interpret *p* as for K-S test)

Independence

- Are the data related:
  - positively (increase in one, increase in other)
  - negatively or inversely (increase in one, decrease in other)
- measures: covariance, correlation coefficient

Covariance

- you may recall: *variance* is squared, unsigned average distance from the mean
- *covariance:* "a measure of the 'average' relationship between two variables. Measured as the average cross-product deviation (the cross-product / n-1)"
- cov(*x*,*y*)= $\sum(x_i - \bar{x})(y_i - \bar{y})/N - 1$

Correlation Coefficient

- Covariance heavily dependent on scale
- We can convert the covariance into standard units by standardization (to the standard deviation)
- *r*=$\sum(x_i - \bar{x})(y_i - \bar{y})/N - 1\ (s_x s_y)$

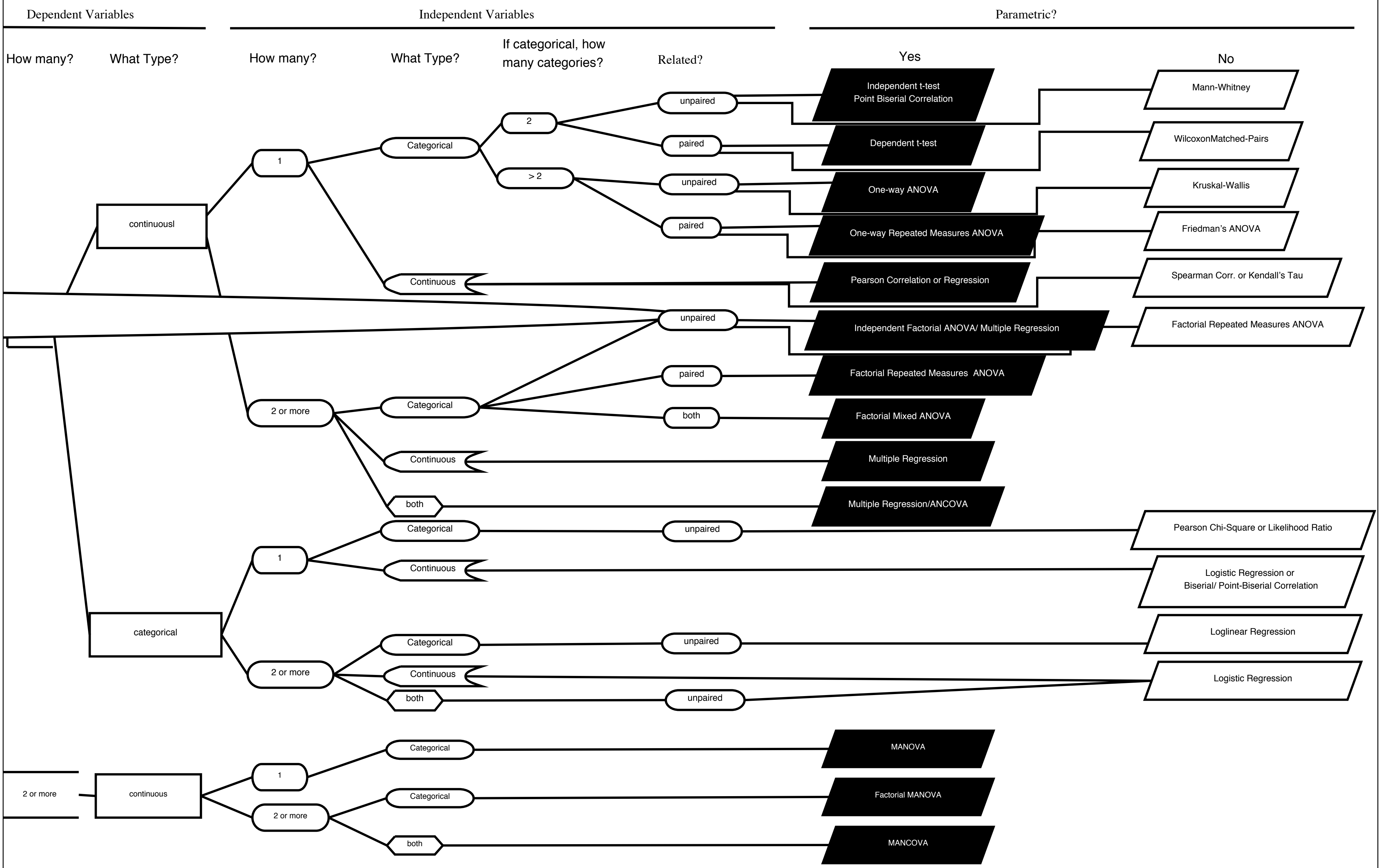
What next?

- Choice #1: correct problems in the data:
  - remove outliers if good reason to believe they're not members of the population being sampled
  - only if appropriate: transform the data (e.g., normalize)
- Choice #2: choose a non-parametric test

Multiple sets of data

- assume sampled from different populations, different means, variances
- use exploratory techniques to see datasets' internal structures
- then compare their structures to each other using same techniques
- use normalized frequency method to examine relative occurrence in your multiple datasets of your variable(s) of interest

- choose an appropriate inferential test


Dependent Variables
Independent Variables
Parametric?
How many?
What Type?
How many?
What Type?
If categorical, how many categories?
Related?
Yes
No
continuousl
1
Categorical
2
unpaired
Independent t-test
Point Biserial Correlation
Mann-Whitney
paired
Dependent t-test
WilcoxonMatched-Pairs
> 2
unpaired
One-way ANOVA
Kruskal-Wallis
paired
One-way Repeated Measures ANOVA
Friedman's ANOVA
Continuous
Pearson Correlation or Regression
Spearman Corr. or Kendall's Tau
unpaired
Independent Factorial ANOVA/ Multiple Regression
Factorial Repeated Measures ANOVA
2 or more
Categorical
paired
Factorial Repeated Measures ANOVA
both
Factorial Mixed ANOVA
Continuous
Multiple Regression
both
Multiple Regression/ANCOVA
categorical
1
Categorical
unpaired
Pearson Chi-Square or Likelihood Ratio
Continuous
Logistic Regression or
Biserial/ Point-Biserial Correlation
2 or more
Categorical
unpaired
Loglinear Regression
Continuous
both
unpaired
Logistic Regression
2 or more
continuous
1
Categorical
MANOVA
2 or more
Categorical
Factorial MANOVA
both
MANCOVA

Lab 8 preview

- Work with your formatted data
- tab-delimited data
- 1 case per row
- Examine R code in file linked to syllabus

References

- Johnson, Keith (2008) *Quantitative Methods in Linguistics*, London: Blackwell