

An Introduction to Statistical Analysis for Linguistics using R

*A practical workshop for Redeemers University Linguistics
Students on the 9th of July 2025*

Temitayo Olatoye
University of Eastern Finland, Finland
temitayo.olatoye@uef.fi

Time	Topic	Description
10:00 – 10:10	Welcome & Overview	<ul style="list-style-type: none"> - Brief introduction - Why R and statistics for linguists
10:10 – 10:35	Basic Statistical Concepts for Linguists	<ul style="list-style-type: none"> - Types of data (nominal, ordinal, interval, ratio) - Descriptive vs. inferential statistics - Variables (e.g., reaction times, word frequency, syntactic choice)
10:35 – 11:05	Getting Started with R & RStudio	<ul style="list-style-type: none"> - R/RStudio basics - R syntax essentials (objects, data frames, indexing) - Loading data (read.csv) and basic inspection: <i>head()</i>, <i>str()</i>
11:05 – 11:40	Exploratory Data Analysis	<ul style="list-style-type: none"> - Summary stats: <i>summary()</i>, <i>mean()</i>, <i>sd()</i>, <i>table()</i> - Visualization with <i>ggplot2</i> (histograms, boxplots, scatterplots) - Detecting trends and outliers
11:40 – 11:50	Break	
11:50 – 12:15	Statistical Tests	<ul style="list-style-type: none"> - t-test <i>t.test()</i> – e.g., comparing mean word durations between groups - Chi-square <i>chisq.test()</i> – e.g., POS vs. speaker group - Correlation <i>cor.test()</i> – e.g., lexical frequency vs. response time
12:15 – 12:50	Intro to Regression Analysis	<ul style="list-style-type: none"> - What is regression analysis? - Linear regression <i>lm()</i>: predicting duration from frequency or age group - Interpreting coefficients and p-values - Logistic regression
12:50 – 13:00	Wrap-Up & Q&A	<ul style="list-style-type: none"> - Resources for further learning - Questions and feedback

Which have you used?

HomeInsertDrawPage LayoutFormulasDataReviewView

Aptos Narrow (Bod... 12 A A

B I U

Generate

N12

	A	B	C	D	E	F	G	H
1	Gender							
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								

Target Corpus

Name: ess

Files: 1

Tokens: 4579

ess12txt.TAGGED.txt

KWICPlotFile ViewClusterN-GramCollocateWordKeywordWordcloudChatAI

File Hits 0File Types 844File Tokens 4579File Name ess12txt.TAGGED.txt

1.0_CRD INTRODUCTION_NN1 The_AT0 wastage_NN1 of_PRF metals_NN2 due_PRP to_PRP corrosion_NN1 has_VHZ become_VVN an_AT0 important_AJ0 engineering_NN1 problem_NN1 ._PUN

With_PRP a_AT0 few_DT0 exceptions_NN2 ,_PUN metals_NN2 are_VBB prone_AJ0 to_PRP corrosion_NN1 in_PRP ordinary_AJ0 aqueous_AJ0 environment_NN1 ._PUN

Most_DT0 of_PRF these_DT0 metals_NN2 gain_VVB wide_AJ0 industrial_AJ0 and_CJC domestic_AJ0 applications_NN2 ._PUN

Probably_AV0 ,_PUN no_AT0 other_AJ0 source_NN1 of_PRF waste_NN1 ._PUN

except_CJS that_CJS affecting_VVG human_AJ0 life_NN1 ,_PUN is_VBZ of_PRF greater_AJC concern_NN1 to_PRP all_DT0 ._PUN

According_PRP to_PRP Speller_NP0 (PUL 1935_CRD)_PUR ,_PUN it_PNP is_VBZ only_AV0 through_PRP the_AT0 elimination_NN1 of_PRF waste_NN1 and_CJC the_AT0 increase_NN1 in_PRP our_DPS national_AJ0 efficiency_NN1 that_CJT we_PNP

can_VM0 hope_VVI to_TO0 lower_VVI the_AT0 cost_NN1 of_PRF living_NN1 ,_PUN

on_PRP the_AT0 one_CRD hand_NN1 ,_PUN and_CJC raise_VVB our_DPS standards_NN2 of_PRF living_NN1 ,_PUN on_PRP the_AT0 other_AJ0 ._PUN

The_AT0 elimination_NN1 of_PRF waste_NN1 is_VBZ a_AT0 total_AJ0 asset_NN1

,_PUN it_PNP has_VHZ no_AT0 liabilities_NN2 ._PUN

An_AT0 accurate_AJ0 estimate_NN1 of_PRF the_AT0 loss_NN1 resulting_VVG

from_PRP corrosion_NN1 of_PRF metals_NN2 is_VBZ out_PRP of_PRP question_NN1 ._PUN

From_PRP certain_AJ0 data_NN0 ,_PUN however_AV0 ,_PUN which_DTQ are_VBB at_PRP hand_NN1 regarding_PRP the_AT0 average_AJ0 annual_AJ0 renewal_NN1 of_PRF corrugated_AJ0 metals_NN2 ,_PUN replacing_VVG corroded_AJ0 parts_NN2 and_CJC preventing_VVG corrosion_NN1 is_VBZ estimated_VVN to_TO0 cost_VVI billions_CRD of_PRF dollars_NN2 in_PRP each_DT0 of_PRF the_AT0 industrialised_AJ0 countries_NN2 ._PUN

The_AT0 cost_NN1 of_PRF corrosion_NN1 to_PRP the_AT0 U.S._NP0 economy_NN1 alone_AV0 is_VBZ estimated_VVN to_TO0 be_VBI \$276_NN0 billion_CRD per_PRP year_NN1 (PUL NACE_NP0 ,_PUN 2002_CRD)_PUR ._PUN

This_DT0 cost_NN1 is_VBZ expected_VVN to_TO0 increase_VVI with_PRP advancement_NN1 in_PRP technology_NN1 and_CJC metals_NN2 utilisation_NN1 ._PUN

The_AT0 greater_AJC percentage_NN1 of_PRF pipelines_NN2 failures_NN2 in_PRP the_AT0 petroleum_NN1 and_CJC gas_NN1 industries_NN2 is_VBZ believed_VVN to_TO0 be_VBI due_PRP to_PRP corrosion_NN1 ._PUN

For_AV0 example_AV0 ,_PUN Nwilo_NP0 and_CJC Badejo_NP0 (PUL 2001_CRD)_PUR reported_VVD that_CJT 50%_NN0 of_PRF the_AT0 oil_NN1 spills_VVZ in_PRP Nigeria_NP0 is_VBZ as_PRP a_AT0 result_NN1 of_PRF corrosion_NN1 of_PRF pipelines_NN2 and_CJC tankers_NN2 ._PUN

Also_AV0 . PUN AFX NP0 Europe NP0 (PUL 2004_CRD) PUR reported_VVD that_CJT

Search Query Words Case Regex Hit Location 0

Start

Adv Search

Progress

*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor									
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help									
13 : InterviewID 13K									
	InterviewID	Name	Gender	Age	Rice	var	var	var	var
1	1	Peter	Male	15.00	Yes				













Introduction – *why R?*

- Open-source
- A growing adoption in linguistics
- Sharing is easy + transparent
- Produce better research
- Do cool stuff... 🤓

RStudio is an Integrative
Development Environment (IDE).



Types of data

Data Type	Ordered?	Equal Intervals?	True Zero?	Example
Nominal (categorical)				Lexical variants, morphosyntactic choices
Ordinal				Evaluation scales, proficiency levels
Interval				Temperature (°C)
Ratio				Reaction time, Frequency

What can you do with your data?

- Description
- Explanation
- Prediction



Descriptive Statistics - Describe a data sample



- Location: Mean value, median, mode, sum



- Dispersion: Standard deviation, variance, range



- Tables: Absolute, relative and cumulative Frequencies



- Charts: Histograms, bar charts, box plots, scatter plots

Inferential Statistics

Make a statement about the general population

Simple test procedures

- t-Test
- Binominal Test
- Chi-square test
- Mann-Whitney U Test
- Wilcoxon-Test
- ...

Correlation analysis

- Pearson Correlation analysis
- Spearman Rank Correlation
- ...

Regression Analysis

- Simple linear regression
- Multiple regression
- Logistic regression
- ...

ANOVA

- Single factorial ANOVA
- Two factorial ANOVA
- ANOVA with measurement repetitions
- ...

Operationalizing your variables

Independent Variable (IV)

Dependent Variable (DV)

What is it?

The variable you **manipulate or group by**

The variable you **measure or observe**

Role in study

Acts as the **cause**, condition, or factor

Acts as the **effect** or outcome

Control over it?



Yes (you choose or categorize it)



No (you observe how it changes)

Examples in Linguistics

- Word Frequency
- Sentence Type
- Speaker Age

- Reaction Time
- Word Duration
- Accuracy

Research question

Does **word frequency** affect **reaction time**?

DV: reaction time (measured in ms)

Analysis

	Monofactorial Analysis	Multifactorial Analysis
Definition	Examines the effect of one independent variable	Examines the effects of multiple independent variables
Complexity	Simpler, easier to interpret	More complex, captures interactions between variables
Example	Studying how age affects pronunciation accuracy	Studying how age, gender, and L1 background affect pronunciation accuracy
Statistical Methods	t-tests, chi-square tests, simple regression	ANOVA, multiple regression, mixed-effects models
Limitations	May overlook confounding factors	Provides a more nuanced and realistic analysis
Usage	Preliminary studies or when data is limited	Advanced research with rich datasets

Resources

- <https://www.stgries.info/research/overview-research.html>
- <http://www.martinschweinberger.de/blog/resources/>
- <http://www.martinschweinberger.de/blog/presentations-talks/>
- <https://dlf.uzh.ch/openbooks/statisticsforlinguists/>
- <https://www.zora.uzh.ch/id/eprint/183632/1/Statistics-for-Linguists-1580118836.pdf>
- https://appliedstatisticsforlinguists.org/bwinter_stats_proofs.pdf

Texts

- Baayen, R Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using r*. Cambridge: Cambridge University press.
- Crawley, Michael J. 2005. *Statistics: An Introduction Using r*. 2005. Chichester, West Sussex: John Wiley & Sons.
- Field, Andy, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using r*. Vol. 50. Sage. <https://doi.org/https://doi.org/10.5860/choice.50-2114>.
- Gries, Stefan. 2009, 2021. *Statistics for Linguistics Using r: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.
- Levshina, Natalia. 2015. *How to Do Linguistics with r: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company.
- Wilcox, R. 2009. *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Vol. 47. Oxford University Press. <https://doi.org/https://doi.org/10.5860/choice.47-5075>.
- Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using r*. Routledge. <https://doi.org/https://doi.org/10.4324/9781315165547>