Beyond example extraction: Quantitative analysis of the JANES corpus

Maja Miličević

University of Belgrade

Slovenščina na spletu in v novih medijih
Ljubljana, 25 November 2015

Why this tutorial?

A BNC example

What is used more frequently, *tutorial* or *workshop*?

`[lemma = "tutorial"]` $\rightarrow$ f=506, `[lemma = "workshop"]` $\rightarrow$ f=2930

χ^2=1710, df=1, $p<0.001$

A JANES example

What is used more frequently, *delavnica*, *workshop* or *tutorial*?

`[lemma = "delavnica"]` $\rightarrow$ f=7873, `[lemma = "workshop"]` $\rightarrow$ f=78, `[lemma = "tutorial"]` $\rightarrow$ f=303

χ^2=14310, df=2, $p<0.001$

Overview

1. Analysing corpus data: How and why?
   - Introduction to quantitative corpus studies
   - R: Basics, data formats and related issues
2. Describing and visualising corpus data
   - Descriptive statistics
   - Graphs
3. Generalising from corpus data
   - Statistical hypothesis testing
   - Some inferential tests

Roadmap

Theory combined with practical examples:

- Examples to do together (full code available)
- Examples to do on your own (by modifying the code)
- Everything based on data extracted from JANES (tweets)

Where to find the data and code?

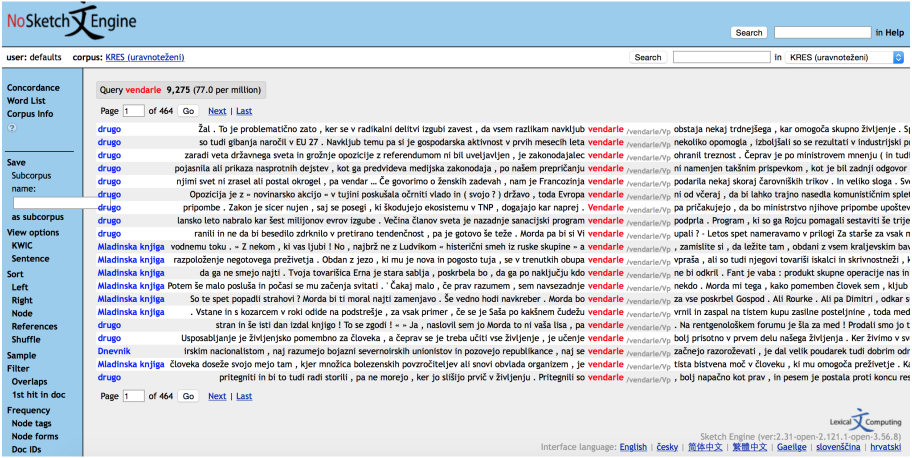
- JANES conference website:
  `http://nl.ijs.si/janes/dogodki/konferenca-2015/`
- Files `tviti.1000.csv`, `tviti.freq.csv`, `code.txt`
  download these files into your R working directory

Overview of topics

1. Analysing corpus data: How and why?
   - Introduction to quantitative corpus studies
   - R: Basics, data formats and related issues

Doing analyses of corpora

http://nl.ijs.si/noske/all.cgi/first_form



Doing more analyses of corpora

Word list

Corpus: **KRES (uravnoteženi)**

Page 1 Go Next >

word	Freq
je	3,289,979
in	2,690,587
v	2,163,634
se	1,523,325
na	1,356,374
da	1,244,624
za	1,086,084
ki	983,036
so	930,592
pa	809,557
z	712,334
ne	610,613
s	548,506
tudi	532,618
bi	529,003
kot	425,159
po	404,952
še	383,543
bo	379,432
ali	374,726

Collocation candidates

Page 1 Go Next >

		Freq	T-score	MI
P | N	ne	304	14.356	2.501
P | N	uspeti	149	12.056	6.339
P | N	lahko	164	10.628	2.556
P | N	priti	119	10.212	3.970
P | N	še	148	9.496	2.188
P | N	morati	108	8.991	2.891
P | N	treba	77	8.261	4.094
P | N	odločiti	67	7.869	4.694
P | N	dobiti	71	7.787	3.720
P | N	najti	67	7.689	4.044
P | N	nekaj	80	7.670	2.812
P | N	nekoliko	61	7.611	5.289
P | N	biti	798	7.570	0.450
P | N	imeti	100	7.116	1.794
P | N	začeti	62	7.030	3.223
P | N	obstajati	49	6.768	4.917
P | N	ostati	50	6.570	3.818
P | N	zgoditi	45	6.327	4.137
P | N	pa	138	6.168	1.074
P | N	iti	54	5.888	2.331
P | N	malce	31	5.415	5.189

Doing quantitative analyses of corpora

Three main kinds/formats of corpus data:

- Concordances
- Frequency lists
- Collocations

All three involve numbers:

- How many times does a certain unit appear in a corpus?
- How are words distributed by frequency?
- How strongly are words associated to each other?

What happens after these initial numbers are obtained?

Additional aspects of quantification

Even though many studies based on data from corpora have a quantitative dimension, many among them do **not** belong to the typical **quantitative research paradigm**

What tends to be missing?

- A hypothesis
- A clear identification and definition of variables
- Inferential statistical tests

Why is this a problem?

While studies missing the above can constitute valuable language descriptions, their findings are **difficult or impossible to generalise**

Quantitative research paradigm

Hypothethical-inferential approach in science:

Start from **theory** → formulate **hypothesis** → develop **methodology** → collect **data** → analyse **data** → accept/refute **hypothesis**

(In many cases **predictions** are tested directly, and hypotheses only indirectly)

Again, this is far from meaning that exploratory studies are not useful - they are, for instance, crucial for developing future hypothesis; the hypothesis formulation process can in fact go both ways

- Cf. the issue of **corpus-based** vs. **corpus-driven** linguistics, i.e. using corpora to *test* vs. to *derive* hypothesis

The role of statistics in linguistic research

In linguistics, there is barely ever access to the **population** researchers are interested in, which can be infinite (Slovene language, nonstandard Slovene language, etc.), so research needs to be based on **samples**

Two kinds of statistics:

- **Descriptive statistics** - describes a sample
- **Inferential statistics** - enables generalisation outside the sample

Samples should (ideally) be:

- **Representative** - a major issue in corpus linguistics
- **Sufficiently large** - small samples are much more prone to non-systematic variation

From research question to generalisation

For inferential statistical tests to be meaningful, what remains central is the **linguistic question** - no statistical analysis will turn a theoretically unsound study into a valid scientific contribution

To maintain theoretical relevance *and* enable generalisation in corpus-based research:

1. Find a question that is **worth exploring**
2. Make sure the question **can be addressed with corpora**
3. Make sure **relevant corpora and relevant data are available**
4. Formulate the question in a way **appropriate for statistical analysis**

Think about how you will analyse the data **before** collecting them!

The importance of research design

How to go about addressing these issues?

1. Find a question that is **worth exploring**
2. Make sure the question **can be addressed with corpora**
   → other types of data more suited? phenomenon too infrequent?
3. Make sure **relevant corpora and relevant data are available**
   → data (in)availability will often influence the research question
4. Formulate the question in a way **appropriate for statistical analysis**
   → research design

Research design

A plan that specifies the **variables** that will be explored, and the ways they will be measured and analysed

Variable types based on the role in the research design:

- **Dependent variable(s)**
  $\rightarrow$ phenomena that we study; a necessary component of research
- **Independent variable(s)**
  $\rightarrow$ factors whose relationship to (impact on?) the dependent variable(s) we want to explore; they can (rarely) be absent (cf. the *workshop* vs. *tutorial* example)

Example: In a study that looks at the use of emoticons in tweets written by female and male users, the use of emoticons is the dependent, and user gender the independent variable

Operationalising variables

Another important notion is that of variable **operationalisation**, i.e. decision on how a variable will be measured

Operationalisation has a theoretical side (what exactly is X – e.g. language standardness – and what is a good indicator of X), but it should also involve statistical (as well as practical) considerations

Many variables can be operationalised in more than one way, depending on the question we want to answer and the kind of data we have available:

- Age: Young, Middle-aged, Old vs. 20, 21, 37, 68, 31...
- Standardness scores (JANES): L1, L2, L3 vs. 1.1, 1.6, 2.7...

The choice of statistical analysis is heavily dependent on how variables are operationalised

Measuring variables (1)

Variable types based on measurement:

- **Qualitative** or **categorical variables**
  gender, part of speech, native language, register, style...
- **Quantitative** or **numerical variables**
  - **Discrete variables** - counted (can only take some values)
    shoe size, number of students, absolute corpus frequencies (number of words/characters/syllables)...
  - **Continuous variables** - measured (can take any value in a range)
    age, height, weight, reaction time in experiments...

Measuring variables (2)

In terms of measurement scales:

- **Nominal scale** → categorical variables
  gender, part of speech, native language, register, style...
- **Ordinal scale** → rank-based variables, intervals between points not necessarily equal (closer to categorical, than to numerical variables)
  Likert scales of the "strongly disagree ... strongly agree" type, army/university ranks, frequency ranks...
- **Interval scale** → numerical variables with an arbitrary 0 and equal intervals between points, but without meaningful ratios
  temperature (C, F, R; 0°C=32°F, 10°C=50°F, 20°C=68°F), IQ...
- **Ratio scale** → numerical variables with a meaningful 0 (=absence) and meaningful ratios
  temperature (K), time, corpus frequencies...

Further clarifications

An additional distinction relevant for categorical data:

- **Binary variables** → two possible values
  yes/no, true/false, active/passive, male/female, corporate/private...
- **Non-binary variables** → more than two possible values
  true/false/not given, positive/neutral/negative, L1/L2/L3...

Interval and ratio scales are treated equally in most statistical analyses

Research designs

Some common research designs by variable type(s):

- **Frequency-based designs** $\rightarrow$ one or more categorical variables
- **Correlational designs** $\rightarrow$ at least two numerical variables
- **Variance-based designs** $\rightarrow$ categorical independent, numerical dependent variable(s) - at least one of each

This list is by no means exhaustive!

Choosing the design

Different designs are possible with the same variables, depending on how they are operationalised; e.g. for the standardness scores in JANES:

- Ordinal scores (L1, L2, L3; T1, T2, T3) → a lot like nominal scores (few values), can be used as categorical (independent) variables in variance- and frequency-based designs, possibly as ordinal (dependent) variables in correlational and variance-based designs
- Interval scores (1.1, 1.6, 2.7...) → can be used in correlational designs, or as dependent variables in variance-based designs

Research designs and statistical tests

Typical **design** → **test mappings**:

- Frequency-based → Chi-square test(s)
  categorical variables, analysing frequencies of mutually exclusive categories
- Correlational → correlation test
  numerical variables, analysing the extent to which they co-vary
- Variance-based (two samples) → Wilcoxon rank sum test / *t*-test
  categorical independent and numerical dependent variable, analysing whether two (sub)categories differ from each other
- Variance-based (more than two samples) → Kruskal-Wallis test / one-way ANOVA
  analysing whether more than two (sub)categories differ

Some particularities of corpora

Some factors that can make corpus data difficult to analyse:

- Corpora as wholes or as collections of individual texts?
  - Analysis by texts usually more reliable (variation between texts taken into account)
  - Availability of data on/from individual texts? Very short individual texts (such as tweets)?
- Absolute or relative frequencies?
  - Absolute fine for same-sized samples
  - Percentages or normalised frequencies (e.g. pmw) otherwise
- Tokens or types?
  - Depends on the research question
  - 'Type' can have different meanings (words vs. constructions)
- How are corpus data distributed?
  - Often not in compliance with statistical desiderata

R

https://www.r-project.org

Data formats

Two things to be careful about:

- **File formats**
  - .csv vs. .txt files
  - Comma-delimited (`read.csv`) vs. tab-delimited (`read.delim`)
  - Use `read.csv2` for semi-colon delimited files, `read.delim2` for tab-delimited files with comma as the decimal mark (important for operating systems set to Slovene regional settings)
- **Character encoding**
  - Excel does not handle diacritics well when saving to .csv or .txt; if your data files contain funny characters (č, š, ž...), the easiest thing to do is to use (freely avalilable) LibreOffice Calc for saving tables

What will we work on?

Two data sets:

- `tviti.freq.csv` - a set of tables containing frequency data extracted from the Twitter subcorpus of JANES, based on available metadata (more info about the (sub)corpus: http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-04-Razvoj-korpusa.pdf)
- `tviti.1000.csv` - various metadata and some other data (numeric standardness scores, number of words, number of emoticons, number of punctuation symbols, etc.) for a random sample of 1000 tweets

What they will look like in R:

- `tviti.freq.csv` will be used to create **crosstabs**
- `tviti.1000.csv` will be imported into R as a **data frame**

The commands you will need are provided in the file `code.txt`

Crosstabs

The example below shows the number of tweets marked 1, 2 and 3 for linguistic standardness, separately for female and male users - the tweets are **cross-classified** by gender and linguistic standardness

	L1	L2	L3
female	782005	268733	100562
male	1748410	506605	144350

The data frame

A portion of the data frame we will create and use is shown below

	id	favorited	retweeted	standard_tech	standard_tech_n	standard_ling	standard_ling_n
1	tid.349449846106759168	0	0	T1	1.1	L1	1.2
2	tid.362949301527252992	1	3	T1	1.1	L1	1.2
3	tid.389334386929201152	0	0	T1	1.2	L1	1.0
4	tid.399105094555144192	0	0	T1	1.2	L1	1.4
5	tid.483870924664737792	1	0	T1	1.2	L2	1.8
6	tid.292193119292780544	0	0	T3	2.9	L1	1.6

Note the following:

- Each column shows one variable
- You can see from the values if a variable is categorical or numerical
- The different values are different **variable levels**
- The first row contains the header with variable names
- All data on one row is about one and the same tweet

Overview of topics

2. Describing and visualising corpus data
   - Descriptive statistics
   - Graphs

What is descriptive statistics used for?

Descriptive statistical measures are important:

- As summary information about the data in the sample
- As starting points for inferential analysis

Typically, descriptive statistical measures show:

- The grouping of data around some value
  → **measures of central tendency**
- The overall distribution of different values in the data
  → **positional measures**
- The dispersion of data around that value
  → **measures of dispersion**

Graphs can also be seen as part of descriptive statistical analysis, and there is no better way of getting an idea of what a data set is like than looking at its graphical representation

Measures of central tendency

Two main measures:

- **Arithmetic mean**
  - the result of averaging the values of individual data points
  - should only be used for interval and ratio scale data
  - very sensitive to extreme values
- **Median**
  - the central value in an ordered list of data values
  - splits the data in two halves
  - 1, 1, 2, 3, 4, 5, 7 $\rightarrow$ 3 // 1, 1, 2, 3, 4, 5, 7, 11 $\rightarrow$ 3.5
  - obligatory for ordinal, can be used for interval/ratio data
  - more robust to extreme values than the mean

Positional measures

Measures of central tendency are also referred to as "location" measures, but not all positional measures reflect a central value:

- **Percentiles**
  - values that split an ordered set of data values into 100 points
- **Quartiles**
  - values that split an ordered set of data values into quarters
  - Q1=25th percentile, Q2=50th percentile, Q3=75th percentile
  - 1, 1, 2, 3, 4, 5, 7, 11, 12 $\rightarrow$ Q1=1.5, Q2=4 (=median), Q3=9 (this is just one possible methods for calculating quartiles, R implements 9 different ones)

Measures of dispersion

Three main measures:

- **Variance**
  - the average of squared differences from the mean
  - should only be used for interval/ratio data
- **Standard deviation**
  - square root of variance, easier to interpret
  - can be used to define outliers
    (data points 2 or more SDs from the mean)
- **Interquartile range (IQR)**
  - the difference between Q3 and Q1
  - can be used to define outliers
    (data points below Q1-1.5*IQR and above Q3+1.5*IQR)

Obtaining descriptive measures in R

Function	Result
`mean()`	mean
`median()`	median
`var()`	variance
`sd()`	standard deviation
`IQR()`	interquartile range
`max()`	maximum value
`max()`	minimum value
`quantile()`	0% (min), 25% (1st quartile), 50% (median), 75% (3rd quartile), 100% (max)
`summary()`	mean, median, 1st and 3rd quartiles, min, max
`sum()` `table()`	frequency counts

`stat.desc()`	available via the `pastecs` package
`count()`	available via the `plyr` package

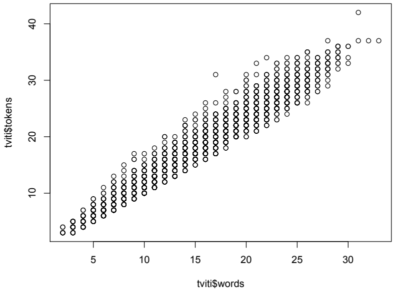
Graph types

Different kinds of graphs are appropriate for different types of data:

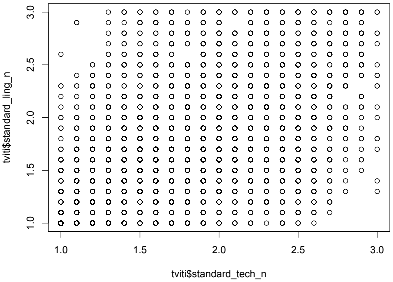
- The well-known types:
  - **Scatterplots** - two numeric variables, individual data
  - **Bar charts** - for summary values, usually grouped by some factor
  - **Line charts** - for summary values, usually grouped by some factor; can also be useful for seeing trends in individual data
- Somewhat less-known types (in linguistic research):
  - **Histograms** - for data distributions
  - **Mosaic plots** - for frequencies by categories
  - **Boxplots** - for data distributions, with a focus on central tendencies

Some simple examples - scatterplots

Number of tokens vs. number of words

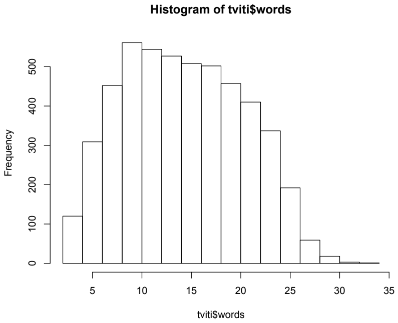


Technical vs. linguistic standardness

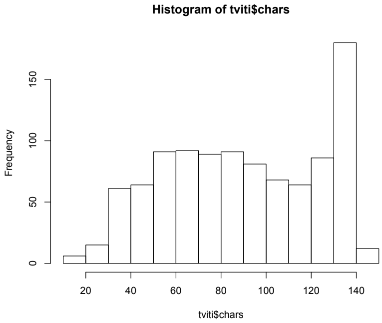


Some simple examples - histograms

Distribution of numbers of tokens



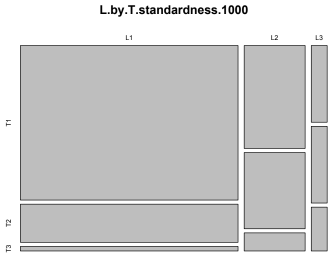
Distribution of numbers of characters



Some simple examples - mosaic plots

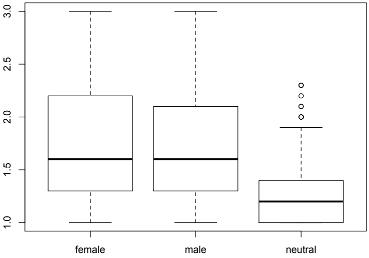
Tweet sentiment by user gender

Linguistic by technical standardness

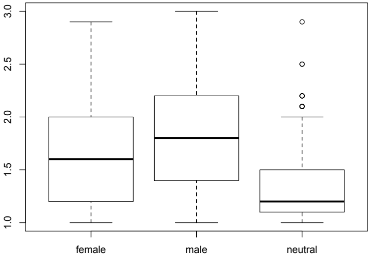


Some simple examples - boxplots

Linguistic standardness by user gender



Technical standardness by user gender



Doing graphs in R

How to do graphs in R?

- Base options cover all of the above graph types
- Specialised packages (e.g. `lattice`, `ggplot2`, and others)

Graphs open in a separate window in R console; from that window, they can be saved in different formats (pdf, png, jpg...)

Note that creating a new graph overwrites the previous graph; to avoid overwriting, open a new graph window before creating a new graph:

Windows - `windows()`
Mac OS X - `quartz()`

Main plotting commands

The above six graph types are covered by five functions:

Function	Graph type
`plot()`	scatterplot, line chart
`barplot()`	bar chart
`hist()`	histogram
`mosaicplot()`	mosaic plot
`boxplot()`	boxplot

Options that can be specified are numerous and varied, ranging from graph titles and bar colours, to the range of information about the data that can be added; R uses default options if not instructed otherwise, but it can be instructed to change pretty much anything

Overview of topics

3. Generalising from corpus data
   - Statistical hypothesis testing
   - Some inferential tests

Back to where we began

What do these numbers mean?

A BNC example

What is used more frequently, *tutorial* or *workshop*?

`[lemma = "tutorial"]` $\rightarrow$ f=506, `[lemma = "workshop"]` $\rightarrow$ f=2930

χ^2=1710, df=1, p<0.001

A JANES example

What is used more frequently, *delavnica*, *workshop* or *tutorial*?

`[lemma = "delavnica"]` $\rightarrow$ f=7873, `[lemma = "workshop"]` $\rightarrow$ f=78,
`[lemma = "tutorial"]` $\rightarrow$ f=303

χ^2=14310, df=2, p<0.001

Statistical significance

Generalisability of sample findings is commonly related to the notion of **statistical significance** of results

Statistical significance is in turn related to **probability**, which can simply be defined as the chance of an event occurring; probability ranges from 0 to 1, i.e. from 0% to 100%

A statistically significant test result indicates a low probability of a research result being due to chance; information on significance is usually expressed as $p>0.05$ or $\mathbf{p<0.05}$, $p<0.01$, $p<0.001$ → these figures stand for non-significant results vs. three different **significance levels** (5%, 1% and 0.1% chance of data being due to chance)

The choice of the 5% level as the cut-off point is a matter of **convention**

What is actually tested?

Two main kinds of hypotheses:

- **Null hypothesis** (H_0)
  assumption that the variables under study are **not** related;
  this result is what would be expected by chance alone
- **Alternative hypothesis** (H_a or H_1)
  research hypothesis (what we actually start from), assumption that the variables under study are related

Even though researchers mostly start their research by assuming a relationship between variables (so some form of H_1), what is actually tested statistically is the H_0

Null Hypothesis Statistical Testing (NHST)

From the perspective of statistics, the only conclusions that can be drawn from an inferential analysis are that **the null hypothesis <u>can</u> or <u>cannot</u> be rejected**

Significance levels obtained through statistical tests show **the probability of the data given the $\mathbf{H}_0$** (not the other way round!)

A word of caution:

NHST is often criticised due to the arbitrariness of the 5% significance level (and some other things); this does not mean that we have just wasted a day, this statistical paradigm is still widely used, just be careful about how you interpret your results!
(see e.g. `https://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/`)

Main results of statistical tests

Typical output from statistical tests contains at least:

- A **test statistic** (χ^2, t, F, W...)
- A **p value**
- Typically, the number of **degrees of freedom**, which are an indication of sample size (or row and column numbers in crosstabs)

These values are reported when writing up the results

In addition, it is highly desirable to calculate and report the **effect size** (https://en.wikipedia.org/wiki/Effect_size), which provides a measure of how meaningful the results are for the given sample size (for large samples significant results can be due to sample size alone)

Parametric vs. nonparametric tests

Two sets of tests that are chosen based on the numerical properties of the data and on their distribution

Parametric tests, which are more powerful, can only be used:

- On interval/ratio scale data
- On normally distributed data
- (On samples with equal variances)

Normality can be checked graphically (e.g. with histograms) and with special tests (e.g. Shapiro-Wilks test); tests also exist for equality of variances (e.g. Levene's or Ansari-Bradley test)

Nonparametric tests make no assumptions about data distributions, and can be also be used on ordinal data; they are calculated on data ranks rather than actual values; they are also better-suited for small samples

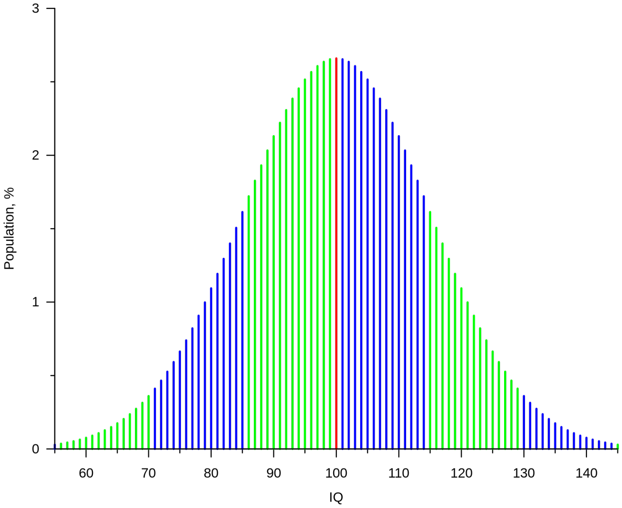
Normal distribution

In populations and sufficiently large samples, a lot of data (of any kind) is distributed **normally**:

- Most population members are grouped around the mean
- Deviations are symmetrical to the mean
    - about 68% of the data falls within ± 1 SD from the mean
    - about 95% of the data falls within ± 2 SD from the mean
    - about 99.7% of the data falls within ± 3 SD from the mean

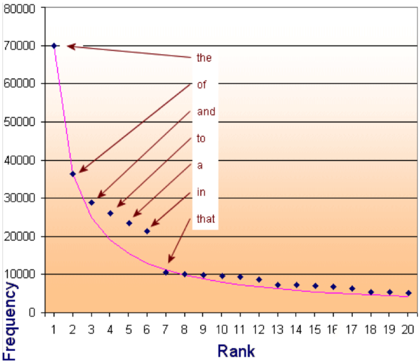
Statistically speaking, this is a highly desirable data distribution;
it is the distribution required for parametric tests

Example - normal distribution of IQ



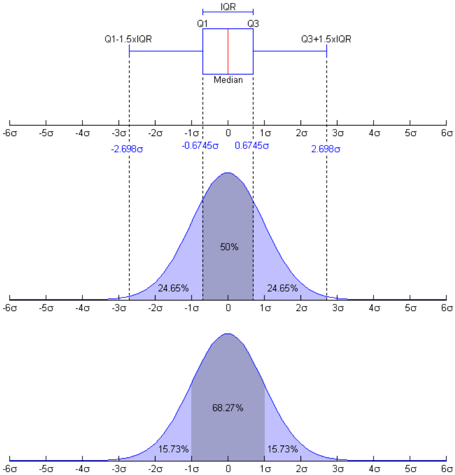
Data distribution in corpora

http://www.intmath.com/exponential-logarithmic-functions/7-graphs-log-semilog.php



Probably more often than not, normal is **not** the kind of distribution found in corpus data, even for non-Zipfian phenomena, so it is important to be aware of nonparametric tests (Fig: Zipfian distribution of word frequencies)

Postional measures vs. the normal distribution curve



Independent vs. dependent samples

Two kinds of samples based on the relationship between what is being compared, important because they require different tests:

- **Independent samples**
  - sample members are independent of each other
  - e.g. tweets by male and female users
- **Dependent samples**
  - sample members are related to each other (paired)
  - e.g. texts from originals and their translations

Not an easy distinction in corpus analysis!
(easier in experimental studies: different vs. same participants)

Research designs, tests, descriptives, graphs

Typical **design** → **test mappings**:

- Frequency-based → Chi-square test(s)
  categorical variables, analysing if these variables are related to or independent from each other; frequencies by category; mosaic plots
- Correlational → correlation test
  numerical variables, analysing whether they co-vary; mean and SD or median and ICQ; scatterplots
- Variance-based (two samples) → Wilcoxon rank sum test / *t*-test
  categorical independent and numerical dependent variable, analysing if the means/medians of two (sub)categories differ from each other; mean and SD or median and ICQ; boxplots (or bar/line plots) [independent samples, version for dependent samples available]
- Variance-based (more than two samples) → Kruskal-Wallis test / one-way ANOVA - means/medians of more than two (sub)categories ...

Doing statistical tests in R

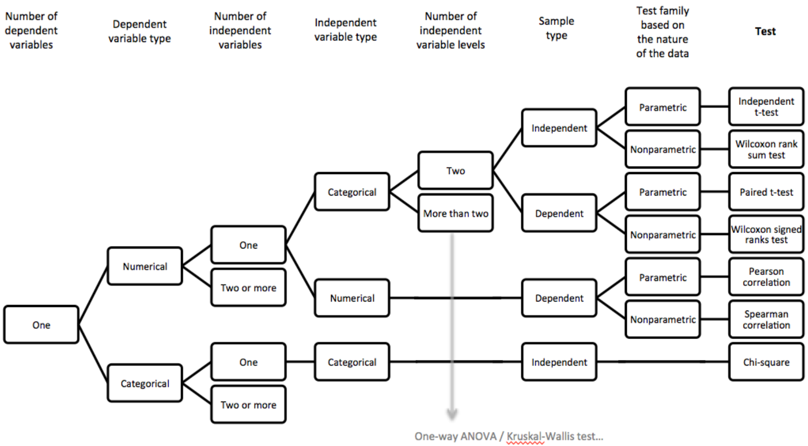
The above tests can be performed via these functions:

Function	**Test**
`chisq.test()`	Chi-square
`cor.test()`	correlation ("spearman"/"pearson", non/parametric)
`wilcox.test()`	Wilcoxon rank sum test (nonparametric)
`t.test()`	*t*-test (parametric)
`kruskal.test()`	Kruskal-Wallis test (nonparametric)
`oneway.test()`	one-way ANOVA (parametric)
`shapiro.test()`	Shapiro-Wilk normality test
`ansari.test()`	Ansari-Bradley equality of variances test

*Add `paired=TRUE` to `t.test()` and `wilcox.test()` to use them as tests for dependent samples

Summary of how to choose a test

For the tests we covered:



Thank you!

m.milicevic@fil.bg.ac.rs

Special thanks to the organisers!

JANES

Univerza *v Ljubljani*
FILOZOFSKA FAKULTETA

SDJT
Slovensko društvo za jezikovne tehnologije

CLARIN.SI