ICAME 2007
Statistics Tutorial
Part 1

Stefan Evert
Cognitive Science, University of Osnabrück
purl.org/stefan.evert

based on joint work with
Marco Baroni, CIMeC, U Trento

sponsored by
UNIVERSITÄT OSNABRÜCK

What is corpus linguistics?

- For me, **corpus linguistics** is concerned with **quantitative statements** about a language
  - quantitative statements = relative frequencies
  - unless you're only interested in the Shakespeare canon, of course ...
- Behind the frequencies is a linguistic question
  - the linguistic phenomenon that we are *really* interested in has to be **operationalised** in terms of relative frequencies for corpus linguistic treatment

Relative frequencies

- Wait a minute, relative frequency means "per million words" (pmw), doesn't it?

Relative frequencies

- Not necessarily ...
  - frequency of *New York City* per million words
  - frequency of noun phrases modified by relative clause as proportion of all noun phrases?
  - frequency of *whom*-relatives as proportion of all places where such a clause could have been used?
    - more examples will be given in part 2
- Criteria
  - corpus as **model of speaker** vs. **model of learner**
  - **familiarity** of phenomenon vs. **choice** probabilities

What is statistics?

- Statistics is about **numbers** … only numbers
  - statistical analysis does not reveal linguistic insights
  - numbers have to be interpreted by the linguist

What is statistics?

- Main task of statistics: draw inferences about a population of objects from a random sample
  - **population** is very large or infinite
  - **objects** have numeric or categorical **properties**
  - statistical methods estimate the **distribution** of such properties from a (small) **random sample**

What is statistics?

- Example: objects are **persons**
  - properties: height, age, shoe size, IQ, … and sex
  - population: all people living in a country, all corpus linguists (past, present and future)
  - distribution: average height, age group proportions, "normal" IQ = 100 ± 30, proportions of men/women
  - sample: all ICAME 2007 delegates (random?)

What is statistics?

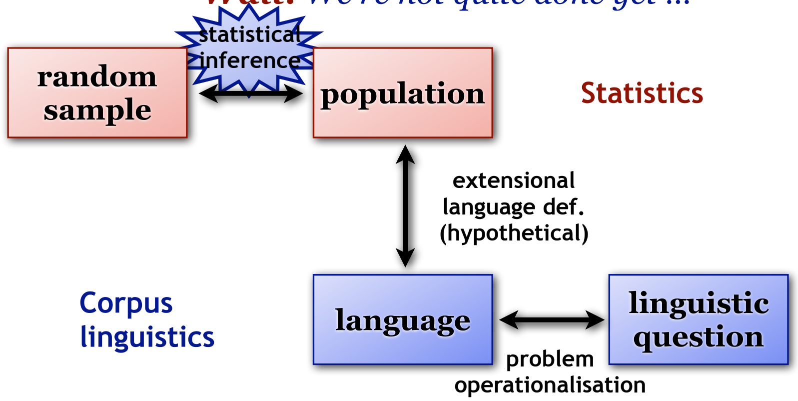
- Example: objects are **noun phrases**
  - properties: length, definite?, adjectives?, subject? (most relevant properties are **binary** = yes/no)
  - population: all noun phrases in a language (or language variety, idiolect, genre, …)
    - refers to noun phrase **tokens**, not noun phrase **types**
    - **extensional definition** of language required for statistics = all existing *and* possible texts in the language
    - unlike in persons example, this is a **hypothetical** population
  - distribution: proportion of definite NPs, subject NPs, …
  - sample: randomly chosen noun phrase tokens **= all noun phrases in a corpus?**

Suspension of disbelief

- We will pretend for now that a corpus is a random sample (of words, NPs, sentences, …)
  - more on this issue in parts 2 and 3

Tutorial summary

Wait! *We're not quite done yet …*



What you will learn now …

- null hypothesis
- p-value & significance
- binomial test
- significance vs. effect size
- confidence interval
- sample size

Toy problem

- American English style guide claims that
  - *"In an average English text, no more than 15% of the sentences are in passive voice. So use the passive sparingly, prefer sentences in active voice."*
  - http://www.ego4u.com/en/business-english/grammar/passive actually states that only 10% of English sentences are passives (as of June 2006)!
- We have doubts and want to verify this claim

Operationalisation

- Problem already phrased in quantitative terms
  - claim: 15% of all sentences are in passive voice
  - side problem: is "passive sentence" a meaningful concept? — we will ignore this issue here
  - passive sentence = contains a passive construction
  - but what is the set of "all sentences"?

Operationalisation

- Extensional definition of a language
  - focus on written, edited American English = *E*
  - language changes over time ➡ synchronic "snapshot"
  - *E* = set of all sentences in all the English books and periodicals published in the U.S. in a certain year
    - the year 1961 is a popular choice …
  - problem: finite set is always incomplete
    - Is *"IBM's new supercomputer has finally beaten the current world chess champion."* not a sentence of English?
  - *E* must include all sentences that *could* have been written ➡ **infinite hypothetical set**

In statistical terms …

- Population = (infinite) set of sentences
  - this is our extensional language definition
- Object = sentence (token)
- Property of interest: *contains passive?*
  - as usual, this is a binary (yes/no) property
- Distribution: proportion of passive sentences
  - we want to find out something about this proportion

Taking a sample

- Cannot count passives in the entire population
  - because it would take far too much time
  - because the population is hypothetical & infinite
- We need to take a **sample** of the population
  - sentences for the sample should be chosen at random
    - 100 sentences from *Rabbit, Run* tell us at best something about how often John Updike uses passive voice
  - sample has to be **representative** of the population
  - good sampling strategy: pick 100 random books from the library, then one random sentence from each book

First results

- 100 sentences in the sample, 19 in passive voice
  - i.e., a proportion of 19%
  - considerably higher than claim of 15%
- Have we falsified the style guide's claim?

Second results

- Let us take another sample just to be sure …
- 100 sentences, 15 in passive voice
  - this is just the claimed proportion of 15%
- Does this sample prove the style guide's claim?

Random variation

- Thought experiment
  - assume that a large number of corpus linguists independently want to verify the style guide's claim
  - each one takes a sample of 100 sentences from the same population
  - (almost) every sample will contain a different number of passive sentences: 19, 15, 21, 26, 14, 22, 17, 25, …
  - some linguists will reject the claim, others not
- **Random variation** introduced by sampling
  - random variation cannot be avoided!

More statistical terminology

- Style guide's claim = **null hypothesis** H_0
  - our goal is to falsify (or reject) the null hypothesis

$$H_0 : \pi = 15\%$$

  - π = proportion of passive sentences in population
- Expected and observed frequency
  - **sample size**: $n = 100$ sentences
  - **expected** frequency: $e = n \times \pi = 15$ passives
  - **observed** frequency: $f = 19$ passives
  - decision based on comparison of f and e

More statistical terminology

- Type I errors (➞ **significance**)
  - assume that null hypothesis is indeed true
  - but we happen to have $f = 19$ passives in our sample
  - unjustified rejection of H_0 ➞ **type I error**
- Type II errors (➞ **power**)
  - assume that H_0 is false, e.g. true proportion $\pi = 19\%$
  - but we happen to find only $f = 16$ passives
  - failure to reject wrong H_0 ➞ **type II error**

Hypothesis tests

- Goal of **statistical hypothesis tests** is to control **risk of type I errors** (false rejection)
- What *is* the risk of a type I error?
  - back to our thought experiment, assuming H_0 is true
  - how many of the corpus linguists would reject H_0?
- Risk of type I error = percentage of random samples for which H_0 would be rejected
  - depends on our rejection criteria, of course!

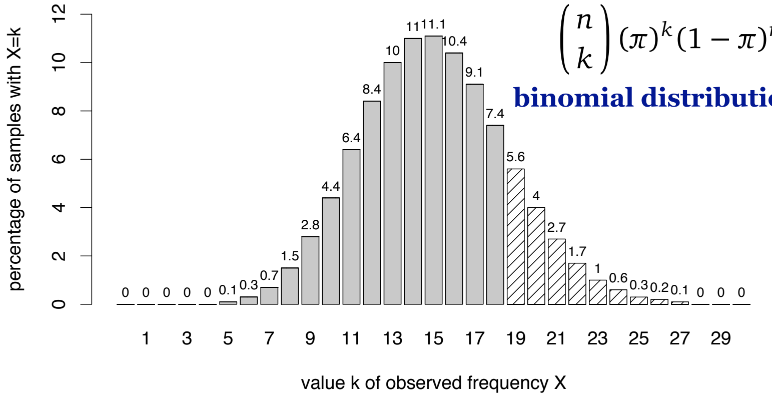
Sampling distribution

- Random variables
  - observed number of passives different in each sample
  - statistical terminology: a **random variable** X
  - X is a "placeholder" for values in different samples, while f is the number observed in a particular sample
- Sampling distribution of X
  - with enough corpus linguists, can tabulate the values of X for many samples ➞ **sampling distribution**
  - perhaps there is a less time-consuming solution?

Sampling distribution

- Random samples = drawing balls from an urn
  - urn with red (= passive) and white (= active) balls
  - proportion of red balls = true proportion π of passives
  - with replacement = sample from infinite population
  - sample frequency X = number of red balls in sample
- With a computer, we don't even need the urn!
  - assume H_0 is true, i.e. 15% of red balls in urn
  - we can now calculate the percentage of samples with a particular passive frequency $X = k$ ($k = 0 \ldots n$)

Sampling distribution



$$\Pr(X = k) = \binom{n}{k} (\pi)^k (1 - \pi)^{n-k}$$

binomial distribution

Sampling distribution

- Probability $\Pr(X=k)$ = percentage of samples for which $X=k$
  - e.g. $\Pr(X=15) = 11.1\%$ of samples have exactly the expected value $e = 15$
  - but $\Pr(X=19) = 5.6\%$ have $f = 19$ ➞ rejection of H_0?
- $\Pr(X=19)$ = risk of false rejection for $f = 19$?

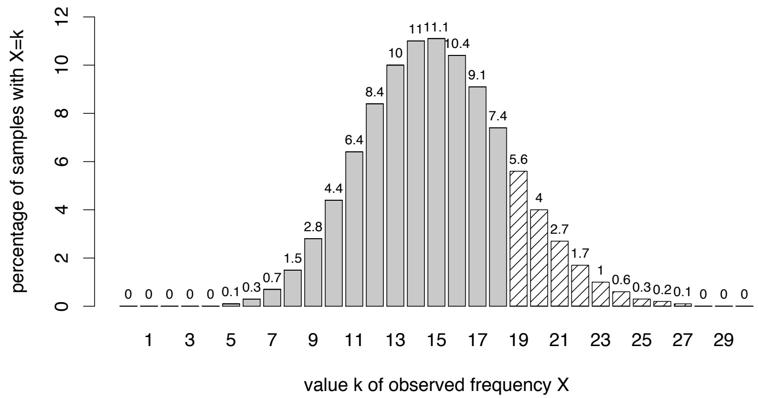
Risk of type I error

- If we are prepared to reject H_0 for $f = 19$, we will also reject it for $f = 20$, $f = 21$, …
- The **risk of a type I error** is therefore:

$$\Pr(X \geq 19) = \Pr(X = 19) + \Pr(X = 20) + \Pr(X = 21) + \cdots + \Pr(X = 100) = 16.3\%$$

- Based on **rejection criterion** $X \geq 19$

Risk of type I error



Congratulations!

- You have just mastered the **binomial test**!
  - choose **rejection criterion**, e.g. $X \geq 19$, based on null hypothesis and expected frequency e
  - calculate **significance** of test = risk of type I error
  - compare observed frequency f to rejection threshold
- Significance level α = “socially acceptable” risk
  - common values are $\alpha = .05$, $\alpha = .01$ and $\alpha = .001$ (i.e. risks of 5%, 1% and 0.1%, respectively)

p-values

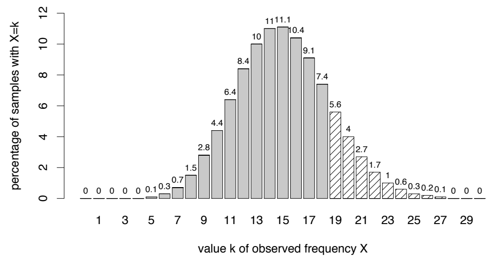
- Q: *“Do I really have to choose a rejection criterion in advance?”*
  - perhaps we could have chosen a much stronger (i.e. more conservative) criterion and still rejected
- A: *In principle, yes. But there's a common practice in statistics …*
  - choose *a posteriori* the most conservative threshold that allows us to reject H_0, i.e. the criterion $X \geq f$
  - type I error $\Pr(X \geq f)$ = **p-value** of the observation f
  - compare p-value with acceptable **significance levels**

One-sided vs. two-sided test

- Our current procedure will only reject H_0 if the true proportion is higher than 15%
  - corresponds to our intuition, i.e. we wanted to disprove the claim “in this direction”
- What if we have no *a priori* expectation?
  - we may just want to test whether the claim is plausible
  - in this case, we should also reject if $f \ll 15$
- What is the correct p-value for $f = 19$ then?

One-sided vs. two-sided test

- In order to calculate **two-sided p-values**, sum over very large and very small values of X
  - include any value X that is “more extreme” than f
  - widely used: chi-squared criterion



$$|X - e| \geq |f - e|$$

or

$$(X - e)^2 \geq (f - e)^2$$

- Q: *"Do I really have to do all this by hand?"*
  - it's going to take ages and I still don't get all the math
- A: *Now that you've understood the principles, you can use statistical software for the math!*
- We (Stefan, Stefan, Harald, …) recommend **R**
  - http://www.r-project.org/
  - more about R in part 3 of the tutorial
  - binomial test: `binom.test(f, n, p=π)`

Significance and effect size

- p-value ➡ significance of evidence against H_0
  - no rejection for $f = 19$ (two-sided p-value $p = .262$)
  - significant rejection for $f = 23$ (two-sided $p = .034$)
  - rejection for $f = 190$ and $n = 1000$ ($p < .001$)
- Significant result ➡ confident that H_0 is wrong
  - but is a **significant** result also **meaningful**?
  - significance is easily achieved for **large samples** (which provide more evidence against H_0)
  - true value of 15.1% not meaningful, but 19% would be

Significance and effect size

- **Significance** tells us whether we have accumulated sufficient evidence to reject
  - boost significance by increasing the sample size
- **Effect size** measures how large the difference between null and true proportion is
  - true effect size (in population) does not depend on the sample, of course
  - but we need large samples to obtain reliable estimates

Effect size & estimation

- In order to measure effect size, we need to estimate the true proportion π of passives and compare it to the null proportion π_0=15%
- Consider a sample with $f = 190$ and $n = 1000$
- The **direct estimate** (**MLE** = maximum-likelihood estimate) for the true proportion is

$$\hat{\pi} = \frac{f}{n} = \frac{190}{1000} = 19\%$$

  - same as for non-significant $f = 19$ and $n = 100$!
  - MLE for true proportion is always unreliable!

Confidence interval

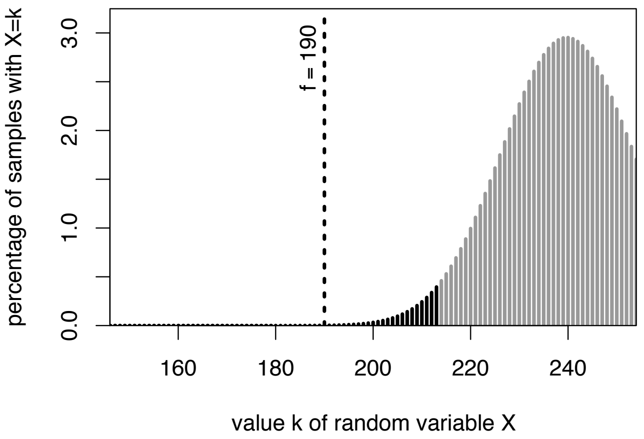
- Goal: estimate a range of plausible values for the true population proportion π
  - based on observed frequency in a sample
  - this range will include the MLE
  - size of confidence interval ➞ "reliability" of MLE
- Set of plausible values = **confidence interval**

Confidence interval

- What is a "plausible value"?
- If we cannot reject H_0: $\pi = 15\%$, then $\pi = 15\%$ is a plausible value
  - i.e. we have no evidence to the contrary
- Use the same logic for other values of π:
  - formulate null hypothesis H_0: $\pi = x$, for any value x between 0% and 100%
  - if binomial test does not reject H_0, percentage x belongs to the confidence interval for π

Confidence interval

$\pi = 24\% \rightarrow H_0$ is rejected



Calculating confidence intervals

- Confidence interval can be computed without testing millions of hypotheses ➞ software
- Size of confidence interval depends on sample size and the significance level of the test

	$n = 100$ $k = 19$	$n = 1{,}000$ $k = 190$	$n = 10{,}000$ $k = 1{,}900$
$\alpha = .05$	11.8%...28.1%	16.6%...21.6%	18.2%...19.8%
$\alpha = .01$	10.1%...31.0%	15.9%...22.4%	18.0%...20.0%
$\alpha = .001$	8.3%...34.5%	15.1%...23.4%	17.7%...20.3%

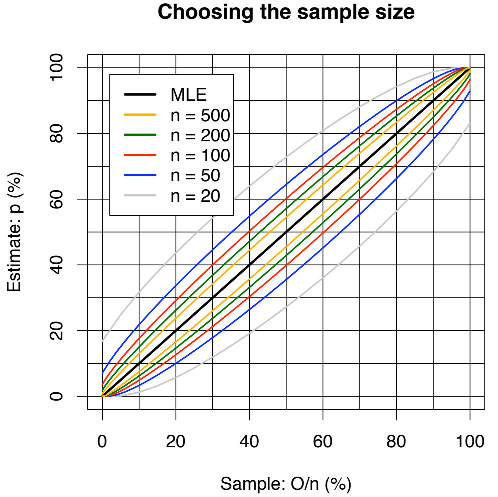
Confidence intervals in R

```
> binom.test(190, 1000, p=.15)

  Exact binomial test

data:  190 and 1000
number of successes = 190,
number of trials = 1000, p-value = 0.0006357
alternative hypothesis: true probability of success is
not equal to 0.15
95 percent confidence interval:
 0.1661265 0.2157137
sample estimates:
probability of success
                  0.19
```

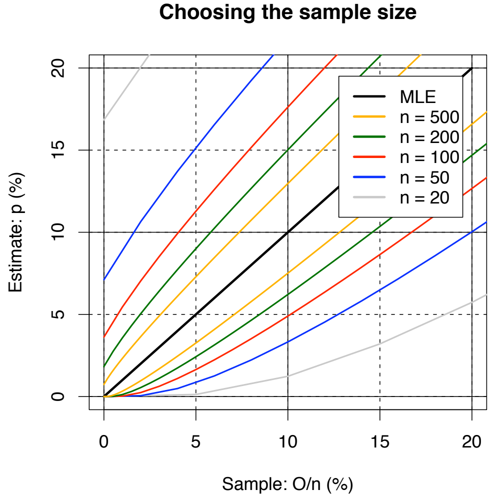

Confidence intervals in R

```
> binom.test(23, 100, p=.15)

  Exact binomial test

data:  23 and 100
number of successes = 23, number of trials = 100, p-
value = 0.03431
alternative hypothesis: true probability of success is
not equal to 0.15
95 percent confidence interval:
 0.1517316 0.3248587
sample estimates:
probability of success
                  0.23
```

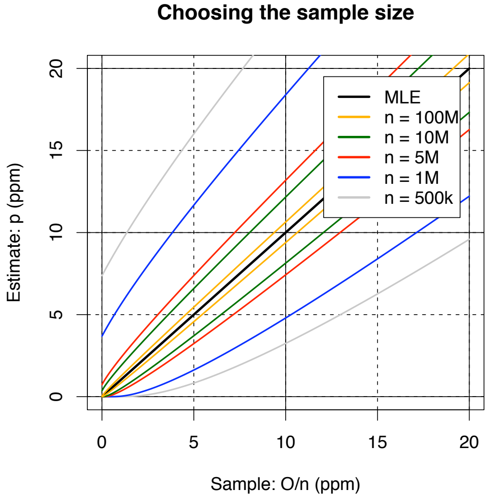

Choosing sample size



Choosing sample size



Choosing sample size



Further reading

- Handout for this part of the course:
  - Baroni, Marco and Evert, Stefan (to appear). Statistical methods for corpus exploitation. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 38. Mouton de Gruyter, Berlin.
  - draft available from purl.org/stefan.evert

Further reading

- Recommended books for further reading
  - Baayen, R. Harald (2007). Analyzing Linguistic Data. A Practical Introduction to Statistics. Cambridge University Press, Cambridge. To appear. http://www.mpi.nl/world/persons/private/baayen/index.html
  - Vasishth, Shravan (2006). The foundations of statistics: A simulation-based approach. Book proposal, University of Potsdam, Potsdam, Germany. http://www.ling.uni-potsdam.de/~vasishth/SFLS.html
  - Butler, Christopher (1985). Statistics in Linguistics. Blackwell, Oxford. http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml