EURAC Seminar on Statistical Methods

Part I

Stefan Evert[†] & Marco Baroni[‡]

[†]Institute of Cognitive Science
Universität Osnabrück
stefan.evert@uos.de

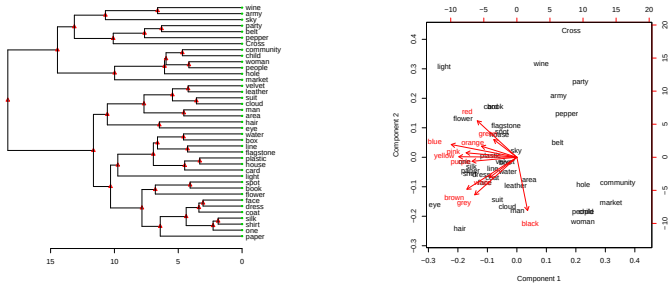
[‡]SSLMIT / SITLEC
Università di Bologna
baroni@sslmit.unibo.it

29 September 2005

A quick warning . . .

With a cookbook of statistical methods and a good software package, you can easily make analyses that . . .

. . . **look good** . . .



. . . are **highly significant** ($p < .001$) . . .

. . . and **completely meaningless**.

What statistics is and what it isn't

Mathematical statistics is quite simple . . .

- it's all about drawing conclusions from a (random) sample

Statistics is just numbers . . .

- it can't reveal linguistic insights by magic
- the numbers need to be interpreted
- use your expert knowledge and *common sense*

You need to ask the right questions . . .

- expand your repertoire of statistical techniques
- ask an expert or a good textbook on statistics
- be clear about what you want to find out
- always ask what the numbers mean
- find out whether your data meet the requirements

Outline

General principles of statistical inference

- (large) **population** of **objects**
  - objects = people, animals, plants, manufactured parts, but also "events" such as the repetitions of an experiment
- objects have measurable **properties** (usually referred to as **random variables**, r.v.)
  - body height, weight, shoe size, age, income, gender, level of education, political opinions, measurements (experiment)
- **scales** of measurement: **nominal** vs. **interval** scale
  - nominal scale = categories (arbitrary labels) ➜ **categorical**
  - interval scale = numbers (meaningful diff's) ➜ **numerical**
- goal: learn something about the distribution of a r.v., characterised by one or more numeric **parameters**
  - average, mode, variance, maximal range, proportion
- either **test hypothesis** about the population or **estimate** a population **parameter**

General principles – some remarks

- random variables are always directly measurable
  - ➥ statistics isn't a magic wand that allows us to discover hidden properties of the objects
- the achievement of statistical analysis is to generalise from measurements on a small number of objects (the sample) to the properties of a large number of objects (the population)
- a sample is a *random* sample, period.
  - ➥ if your sample isn't random, you either have to change the sampling procedure, or you just *pretend* it's random
- the **randomness assumption** has recently become a major issue in (statistically-minded) corpus linguistics

Statistical inference & corpus frequency data

- **population** = language, sublanguage, idiolect, ...
  - e.g. British English, technical writing, specific domain, ...
  - need for observability ➜ **extensional definition** of the population as all text that has ever been written/spoken (or that could be written/spoken) in the language/variety
- **objects** = word, sentences, documents, ...
  - objects are (word) *tokens* rather than *types*
- **properties** of words are usually on nominal scale
  - verb?, noun?, plural noun?, imperative? transitive use?, ...
  - word category, type (word form vs. lemma), word = *torque*?
  - properties marked '?' are **binary** r.v. (1 = yes, 0 = no)
- some random variables on interval scale:
  - word length, no. of syllables, position in sentence, ...
  - more useful properties for sentences and documents

Statistical inference & corpus frequency data

- distribution of nominal properties (esp. binary r.v.) is characterised by **proportions** as parameters
  - What is the proportion of nouns in English? (e.g. 20%)
  - How often does a speaker use nouns? (German > English)
  - How often does the word *the* occur? (ca. 6%)
  - How often does the word *torque* occur? (ca. 5.5 times per million tokens = 5.5 ppm)
  - How often is the verb *give* used without indirect object? (usually measured as proportion of all instances of *give*)
- **sample** = corpus (existing or specially constructed)
  - some corpora are true samples (Brown, BNC)
  - but often the randomness assumption is highly problematic
  - in this course, we will pretend that corpus = random sample

Examples of corpus frequency data

The disturbance of generalized cheerfulness (serene-calmness disorder) is an illness , which is often recognised very late . It 's characterised by a uniformity of psychic experience in the face of circumstances which would otherwise give cause for depressiveness , despair , great anxiety , self-accusations or aggression directed against others . Early diagnosis and specific therapy may decisively improve life-quality of the patients and may also prevent detrimental individual and social long-term consequences .

1. identify tokens (here: words)
2. count tokens: $n = 68$
3. locate nouns
4. count nouns: $k = 21$ ➜ proportion: $\hat{p} = k/n = 30.9\%$
5. locate instances of *of*
6. count instances of *of*: $k = 4$ ➜ $\hat{p} = 5.9\%$

Basic inference: Hypothesis testing

Example of a research question

Are there more nouns in technical writing than in general language (here, in international English)?

- corpus linguists claim that the avg. proportion of nouns in English is 22% (from years of experience with corpora)
- this is our **null hypothesis** $H_0 : p = .22$ (technical writing has the same proportion of nouns as general English)
- also called the **null proportion** p_0 ($H_0 : p = p_0$)
- **population**: all technical writing from the relevant domain
  - here, we are looking at psychology and psychiatry
- **sample**: tokens from a number of available documents
  - we pretend that this is a random sample from the population
- goal: use sample to **test** the null **hypothesis**

Basic inference: Hypothesis testing

- my sample is an abstract from a relevant scientific journal (*Forum der Psychoanalyse*)

The disturbance of generalized cheerfulness (serene-calmness disorder) is an illness, which is often recognised very late. It is characterised by a uniformity of psychic experience in the face of circumstances which would otherwise give cause for depressiveness, despair, great anxiety, self-accusations or aggression directed against others. Early diagnosis and specific therapy may decisively improve life-quality of the patients and may also prevent detrimental individual and social long-term consequences.

- sample size: $n = 68$ word tokens (excluding punctuation)
- if the null hypothesis H_0 is true, the **expected frequency** (i.e. number of nouns in the sample) is $E = n \cdot p \approx 15$
- **observed frequency**: $O = 21 > E = 15$ ➜ reject H_0?
- another perspective: $\hat{p} = 30.9\% > p_0 = 22\%$ ➜ reject H_0?

☞ Why is this conclusion not valid?

Basic inference: Hypothesis testing

- this particular abstract was an arbitrary choice
- what if another abstract had been selected?

$\hat{p} = 30.9\%$	$(21/68)$
$\hat{p} = 25.3\%$	$(49/194)$
$\hat{p} = 21.4\%$	$(21/98)$
$\hat{p} = 23.9\%$	$(26/109)$
⋮	⋮

- observed frequency is subject to random variation

➥ when is the difference $O - E$ (or the difference $\hat{p} - p_0$) large enough to reject H_0 with confidence?

The sampling distribution

- how much random variation do we have to expect?
  - answer depends on true proportion p in the population
- look at many different samples and tabulate their values
  - such additional samples are usually not available
- we can *simulate* these samples if we assume that
  - they are random sets of n tokens from the population
  - the null hypothesis H_0 is true (so we "know" p)
- use different example (for presentation purposes)
  - frequency of the word *the* ($p_0 = 6\%$)
  - all samples have the same size $n = 100$ (easier to tabulate)

UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction

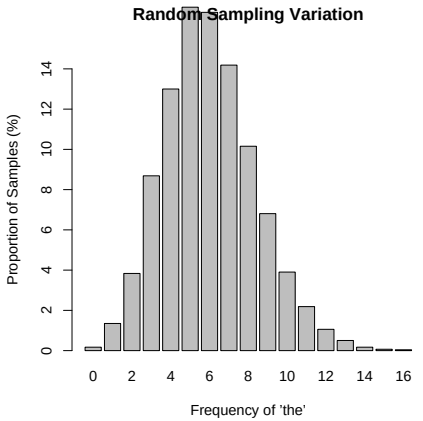
General principles of statistical inference

Statistical inference and corpus frequency data

Basic inference: Hypothesis testing

Basic inference: parameter estimation

The sampling distribution



UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction

General principles of statistical inference

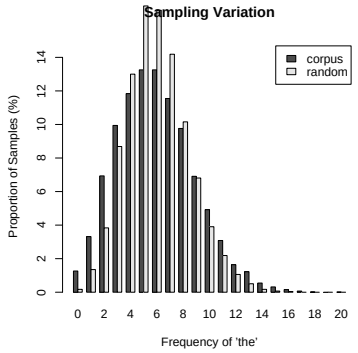
Statistical inference and corpus frequency data

Basic inference: Hypothesis testing

Basic inference: parameter estimation

An aside: the randomness assumption

- testing the **randomness assumption**:
  - compare random sampling distribution with true distribution obtained from a large number of real samples
  - null hypothesis must hold for the sampling population
- sufficient number of samples must be available
  - here: 11,708 samples of 100 tokens each from Brown corpus



UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction

General principles of statistical inference

Statistical inference and corpus frequency data

Basic inference: Hypothesis testing

Basic inference: parameter estimation

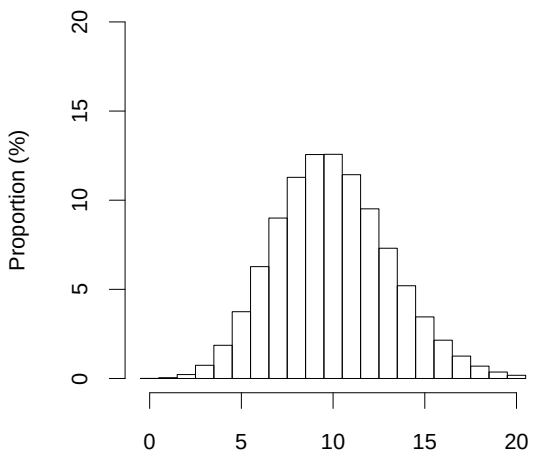
The binomial distribution

- mathematically, the observed frequency of *the* in a sample is a **random variable** F
- we pick a sample with $F = k$ with a **risk** of $\Pr(F = k)$ (statisticians call this risk a **probability**)
- if we know p, we can obtain the probabilities by simulation
- but we can also work them out directly:

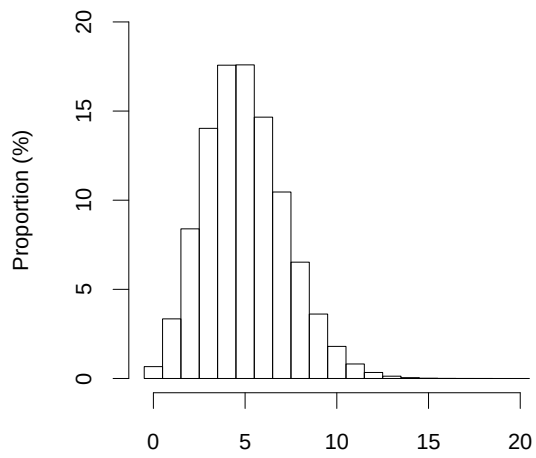
$$\Pr(F = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- **binomial distribution** $F \sim B(n, p)$ with parameters
  - sample size n (determined by sampling process)
  - success probability p (population characteristic = parameter)
- **R** command: `dbinom(k, n, p)`
- long-time average: **expectation** $\mathrm{E}[F] = n \cdot p$

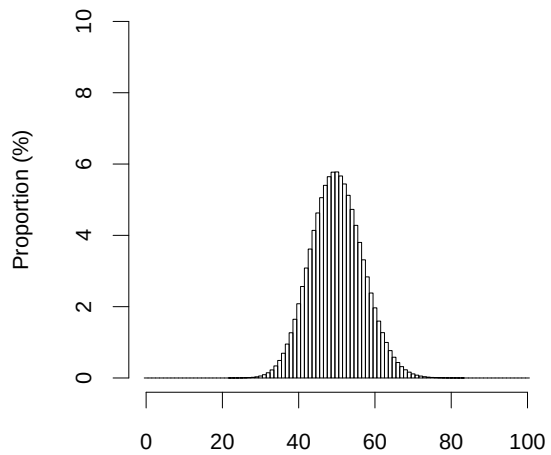
Examples of binomial distributions



Examples of binomial distributions

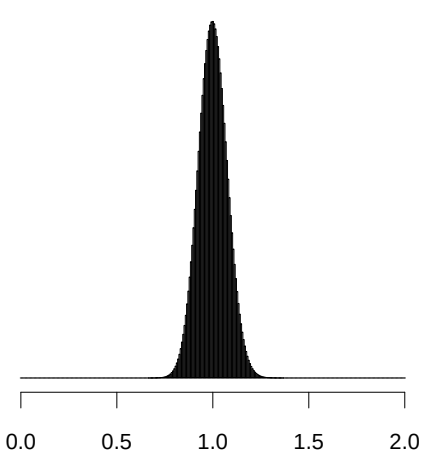


Examples of binomial distributions

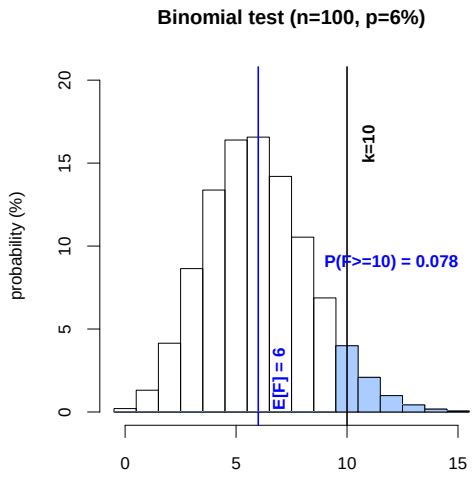


The law of large numbers

Relative sampling error depends on expectation



Testing the binomial hypothesis



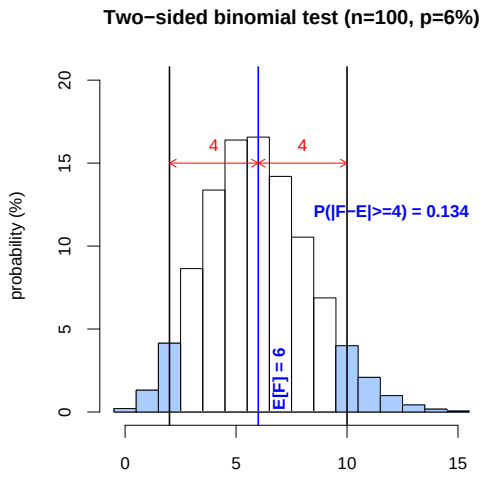
UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction
General principles of statistical inference
Statistical inference and corpus frequency data
Basic inference: Hypothesis testing
Basic inference: parameter estimation

Testing the binomial hypothesis



UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction
General principles of statistical inference
Statistical inference and corpus frequency data
Basic inference: Hypothesis testing
Basic inference: parameter estimation

The binomial test

- reject H_0 if observation O is unlikely given H_0

☞ non-rejection does not count as positive evidence!

- **decision criterion**: deviation $O - E$
  - $O - E$ is an instantiation of $F - \mathrm{E}_0[F]$ for the given sample
  - has to explained by chance if we believe in H_0!
- **one-sided test**: reject H_0 if $O \geq L = E + \delta$
  - "left" one-sided test: reject H_0 if $O \leq L' = E - \delta$
- **two-sided test**: reject H_0 if $|O - E| \geq \delta$
  - this is called a **chi-squared criterion**
  - use two-sided test unless you know exactly what you're doing!
- risk of false rejection (**type I error**, one-sided test)

$$\Pr\big(F \geq L \,|\, H_0\big) = \sum_{k=L}^{n} \binom{n}{k} (p_0)^k (1 - p_0)^{n-k}$$

(similar summation for two-sided test)

UNIVERSITÄT OSNABRÜCK

Statistical Methods Part 1

Evert & Baroni

Introduction
General principles of statistical inference
Statistical inference and corpus frequency data
Basic inference: Hypothesis testing
Basic inference: parameter estimation

The binomial test: significance level

- determine δ (or L) so as to control risk of type I error
- pre-defined **significance level** α (max. acceptable risk)
  - $\alpha = .05 = 5\%$ (95% confidence)
  - $\alpha = .01 = 1\%$ (99% confidence)
  - $\alpha = .001 = .1\%$ (99.9% confidence)
- choose L such that $\Pr\big(F \geq L \,|\, H_0\big) < \alpha$ (one-sided test)
- or δ with $\Pr\big(|F - \mathrm{E}_0[F]| \geq \delta \,|\, H_0\big) < \alpha$ (two-sided test)
- ideally, the type I risk should match α exactly, but this is rarely possible because F has a **discrete** distribution
- better approach: set $L = O$ (or $\delta := |O - E|$) and compute the type I risk corresponding to this threshold ➜ **p-value** (this is often called an **exact test**)
- **R** command: `binom.test(O, n, p0)` ➜ p-value etc.
- two-sided test, use option `alternative="greater"` for one-sided test (or `alternative="less"` for "left" test)

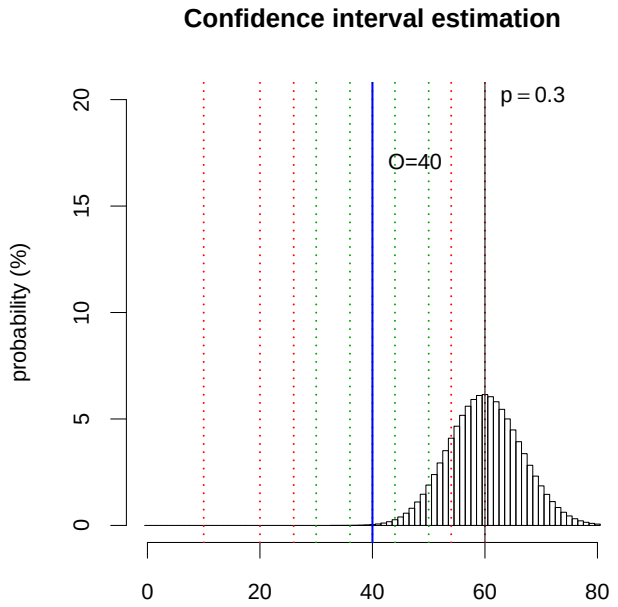
Estimating parameters

- how can we estimate the parameter p without having to formulate a hypothesis or make a guess?
- statistics is a *what if?* game
  - try all possible null hypotheses $H_0 : p = p_0$
  - reject p_0 if sample provides evidence against H_0
  - all values of p_0 that were not rejected are plausible estimates for the parameter p

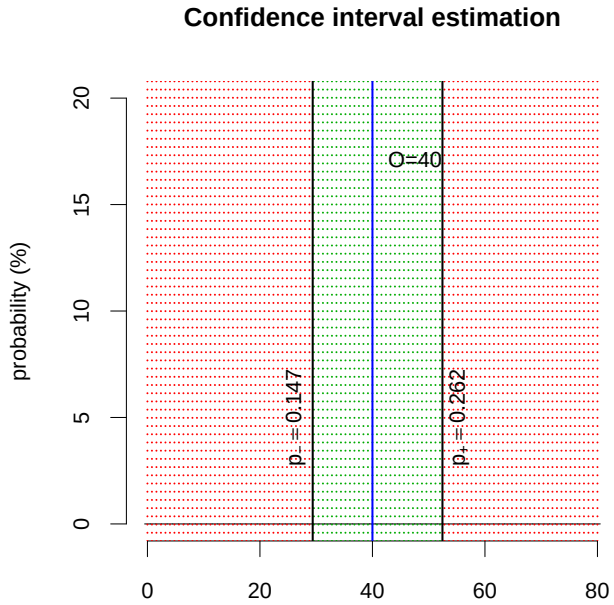
➥ **confidence set** (or **interval**) for p

- depends on pre-defined confidence level $1 - \alpha$
- typical: 95% confidence ($\alpha = .05$) etc.

Binomial confidence intervals



Binomial confidence intervals



Binomial confidence intervals

- binomial confidence intervals can be computed in **R**

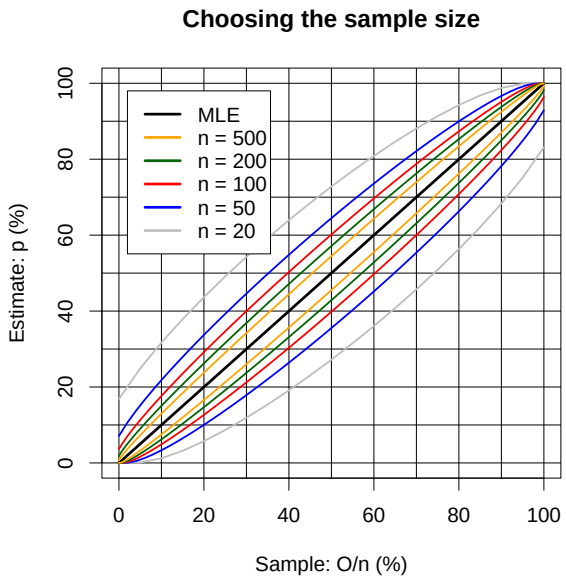
```
result <- binom.test(O, n, conf.level=.95)
p.lower <- result$conf.int[1]
p.upper <- result$conf.int[2]
```

- avoid numerical problems with large samples by using `prop.cint` from our `corpora` library

```
result <- prop.cint(O, n, conf.level=.95)
p.lower <- result$lower
p.upper <- result$upper
```

☞ O and n can be vectors!

How much data do I need?



UNIVERSITÄT OSNABRÜCK

How much data do I need?

Statistical Methods Part 1

Evert & Baroni

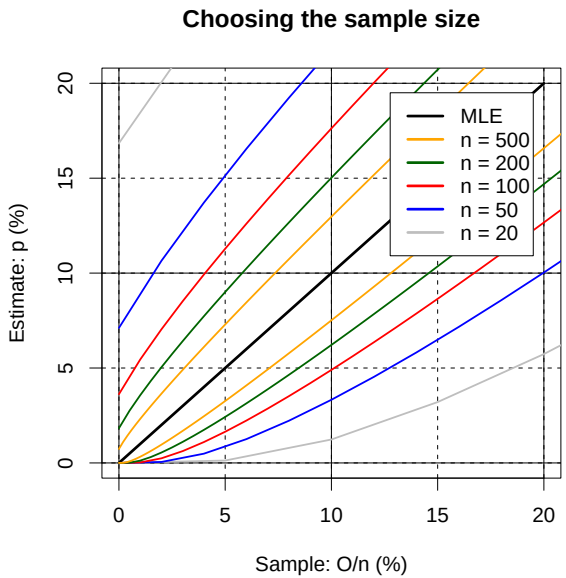
Introduction

General principles of statistical inference

Statistical inference and corpus frequency data

Basic inference: Hypothesis testing

Basic inference: parameter estimation



UNIVERSITÄT OSNABRÜCK

How much data do I need?

Statistical Methods Part 1

Evert & Baroni

Introduction

General principles of statistical inference

Statistical inference and corpus frequency data

Basic inference: Hypothesis testing

Basic inference: parameter estimation