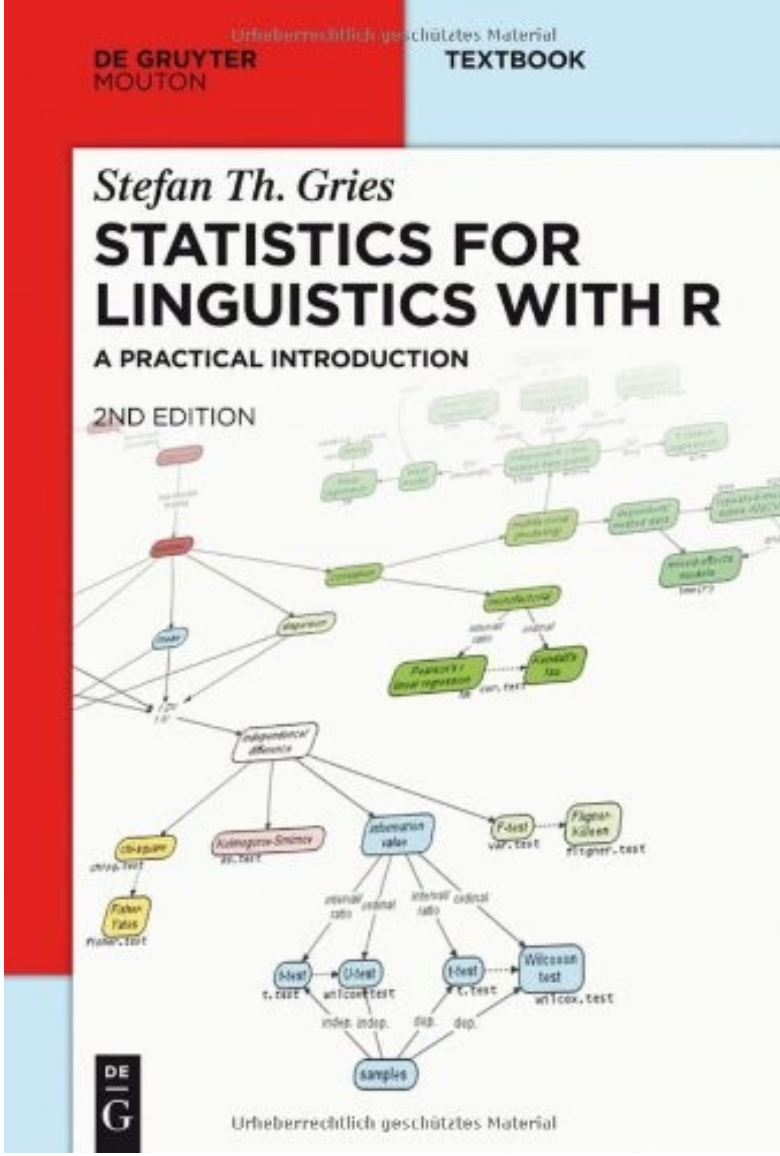
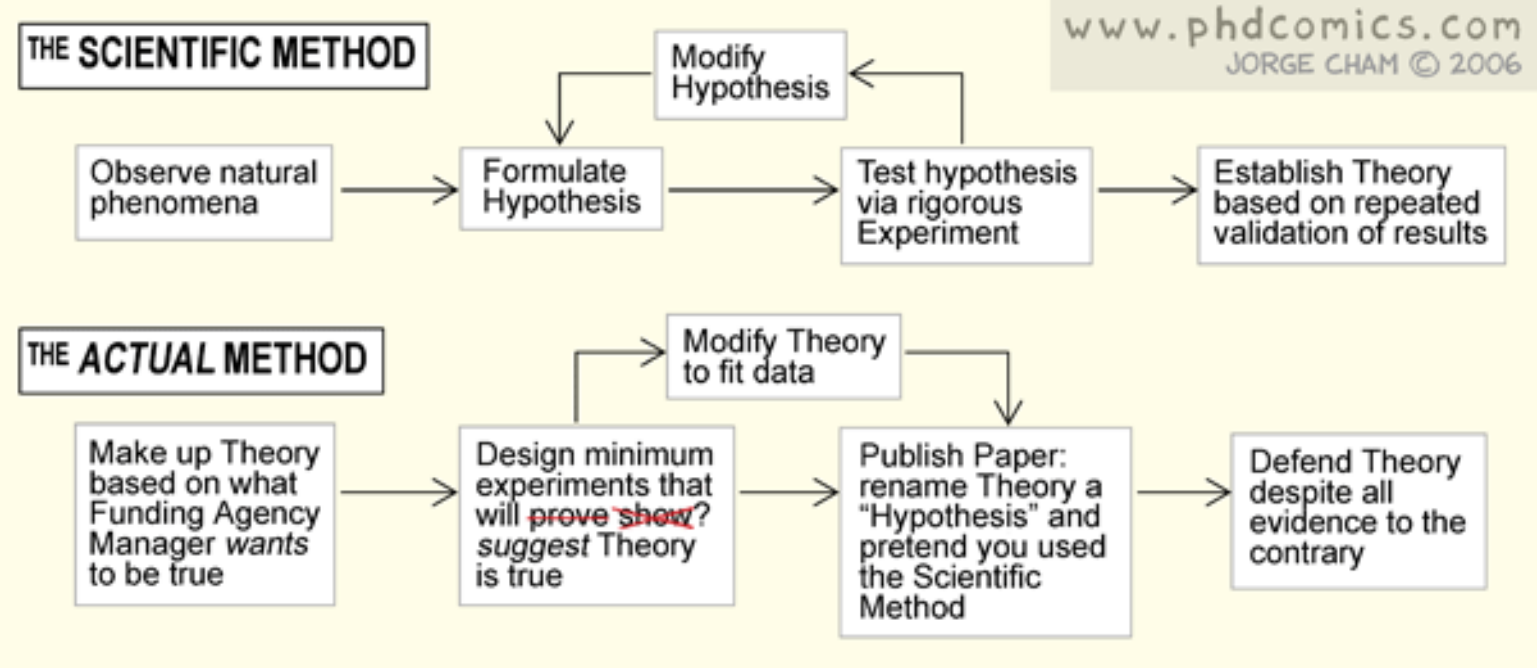
Research methods and statistics for linguists



Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara

A few self-evident characteristics of empirical scientific inquiry



A few self-evident characteristics of empirical scientific inquiry

- Objectives
  - description: what happens?
  - explanation: why does x happen?
  - prediction: what will happen with x if …?
  - control: how can x be influenced?
- why use statistics for this?
  - to describe, explain, predict objectively and comparably
  - to describe, explain, predict precisely and concisely
  - to cope with variability and to generalize
    - one usually doesn't study the population but only a sample
    - different samples will yield different results so we need to
      - quantify this variability
      - separate random from systematic/meaningful variation
  - to assess the reliability of one's generalizations
- central notions
  - objectivity: independence of personal opinions
  - reliability: precision
  - validity: one measures what one wants to measure

Pitfalls to avoid: overgeneralizations

- A published study on two English verbs A and B compared their complementation patterns on the basis of the following data

Verb	Pattern 1	Pattern 2	Totals
A	295 (74%)	104	399
B	131 (79%)	35	166
Totals	429	139	565

- one conclusion drawn from this data was "[c]omparing the postverbal elements in the two verbs, we can see that the proportion of [Pattern 1] for [*B*] is higher than for [*A*]"
- yes, 79% > 74%, but …

Pitfalls to avoid: overgeneralizations

- A published study on two English verbs A and B compared their complementation patterns on the basis of the following data

Verb	Pattern 1	Pattern 2	Totals
A	295 (74%)	104	399
B	131 (79%)	35	166
Totals	429	139	565

- one conclusion drawn from this data was "[c]omparing the postverbal elements in the two verbs, we can see that the proportion of [Pattern 1] for [*B*] is higher than for [*A*]"
- yes, 79% > 74%, but …
- the distribution is not significantly different from chance

```
> freq.table <- matrix(c(295, 131, 104, 35), ncol=2)
> chisq.test(freq.table, correct=FALSE)

 Pearson's Chi-squared test

data:  freq.table
X-squared = 1.5679, df = 1, p-value = 0.2105
```

Pitfalls to avoid: oversight

- A published study on two English expressions *x* and *y* discussed the distribution of different kinds of XPs after *x* and *y*; these were the results

Expression	NP	PP	VP	AdjP	AdvP	Totals
X	302	8	145	19	8	482
Y	73	0	5	3	0	81
Totals	375	8	150	22	8	563

- one conclusion drawn from this data was "[i]f we look at the distribution of *x* before major constituents we find that (a) *X* is more common before noun-phrases than before other constituents"
- yes, 302 = largest figure in the first row, but …

Pitfalls to avoid: oversight

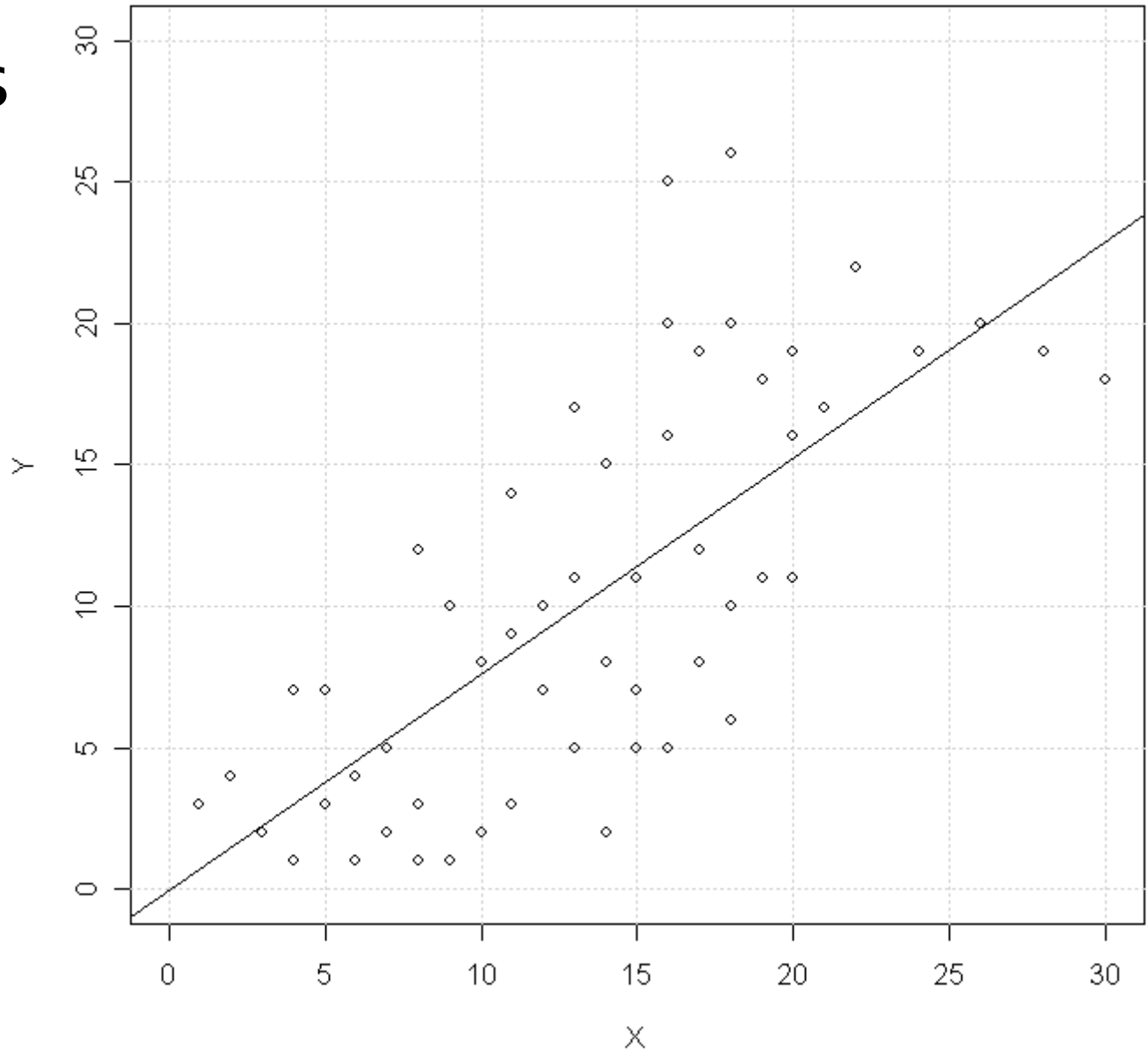
- A published study on two English expressions *x* and *y* discussed the distribution of different kinds of XPs after *x* and *y*; these were the results

Expression	NP	PP	VP	AdjP	AdvP	Totals
X	302	8	145	19	8	482
Y	73	0	5	3	0	81
Totals	375	8	150	22	8	563

- one conclusion drawn from this data was "[i]f we look at the distribution of *x* before major constituents we find that (a) *X* is more common before noun-phrases than before other constituents"
- yes, 302 = largest figure in the first row, but …
- the focus of much of the study was on *x* compared to *y*, and compared to *y*, *x* actually *dis*prefers to occur before NPs

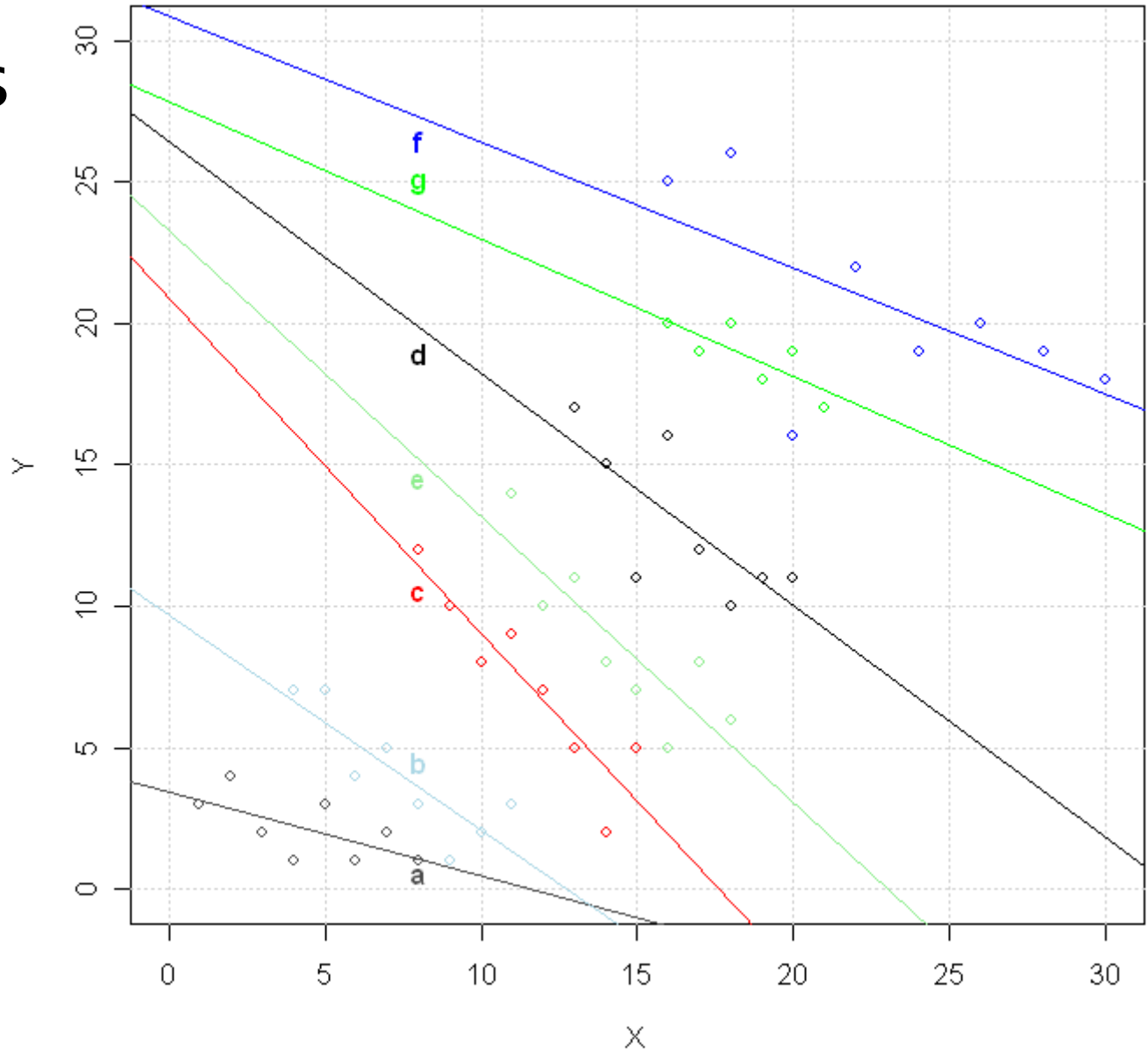
```
> freq.table <- matrix(c(302, 73, 8, 0, 145, 5, 19, 3, 8, 0), ncol=5)
> chisq.test(freq.table, correct=FALSE)$res
          [,1]      [,2]      [,3]       [,4]      [,5]
[1,] -1.063074  0.439797  1.463157  0.0380622  0.439797
[2,]  2.593250 -1.072836 -3.569209 -0.0928485 -1.072836
```

Avoiding false interpretations



(Example from Crawley)

Avoiding false interpretations



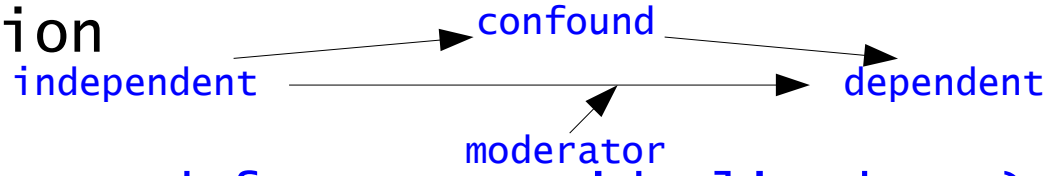
(Example from Crawley)

Caveats regarding, and the structure of empirical quantitative studies

- Note, however,
  - statistics don't provide content – it's always the researcher who does that
  - statistics are only useful to the extent that the researcher has been successful
    - in operationalizing his variables appropriately
    - eliciting/collecting the data correctly
    - choosing the right statistical technique
- phases of an empirical quantitative study
  - reconnaissance
  - formulate hypotheses (text and statistical form)
  - data collection ((operationalizations of) variables)
  - evaluation of hypotheses in the light of the data
    - significance test (*p*-values)
    - effect sizes
    - graphs

Phase (1 and) 2: variables

- Variables are symbols of sets of characteristics an item can exhibit
- they can be distinguished in terms of
  - their role in an investigation
    - independent
    - dependent
    - confounding (controlled, accounted for, or residualized out)
    - moderator (accounted for by interactions w/ add. variables)
  - their information value
    - nominal/categorical
      - different values → different properties
    - ordinal
      - nominal/categorical +
      - different values → different ranks
    - interval/ratio
      - nominal/categorical + ordinal +
      - different values → sizes of differences



Phase (1 and) 2: variables

Time	Rank	Name	Number	Medal
9.86	1	S. Davis	453473	1
9.91	2	J. White	563456	1
10.01	3	S. Hendry	756675	1
20.02	4	C. Lewis	585821	0

Phase (1 and) 2: variables

Time	Rank	Name	Number	Medal
9.86	1	S. Davis	453473	1
9.91	2	J. White	563456	1
10.01	3	S. Hendry	756675	1
20.02	4	C. Lewis	585821	0
int/ord/cat	ord/cat	nom/cat	nom/cat	all

Phase 2: Text hypotheses → operationalization → statistical hypotheses

- What are hypotheses?
  - universal statement (going beyond a singular event)
  - implicit structure of a conditional sentence
    - if …, then …
    - the more/less …, the more/less …
  - empirically testable and potentially falsifiable
  - statements postulating a distribution of (a) dependent variable(s) or statements relating (an) independent variable(s) to (a) dependent variable(s)
- kinds of (text and statistical) hypotheses
  - alternative hypothesis H1: a statement postulating
    - a particular distribution of a (dependent) variable
    - a relation between 2+ independent and/or dependent variables
  - null hypothesis H0: the logical counterpart to H1, an alternative hypothesis with *not* in it

Phase 2: Text hypotheses → operationalization → statistical hypotheses

- Operationalization
  - definition 1
    - pinpointing and fleshing out the notions that the text hypotheses refer to
  - definition 2
    - translating the text hypotheses into something that involves numbers (i.e., can be counted, measured, …)
  - most frequent statistical measures
    - counts/frequencies
    - distributions
    - averages/means
    - dispersions
    - correlations
- examples: how would one operationalize
  - the physical fitness of humans
  - the financial wealth of a person
  - the lengths of sentences
  - the givenness/accessibility of referents of subjects
  - the knowledge of a foreign language

The younger bachelors ate the nice little parrot.

Phase 2: Text hypotheses → operationalization → statistical hypotheses

- On the basis of the operationalization, the text hypotheses H0 and H1 are 'translated' into statistical hypotheses
  - text form
    - H1: subjects are shorter than direct objects
    - H0: subjects are not shorter than direct objects
  - statistical form 1
    - H1: $\text{mean}_{\text{word length of subjects}} < \text{mean}_{\text{word length of dir. objects}}$
    - H0: $\text{mean}_{\text{word length of subjects}} \geq \text{mean}_{\text{word length of dir. objects}}$ (often =)
  - statistical form 2
    - H1: no. of subjects longer than average < no. of dir. objects longer than average
    - H0: no. of subjects longer than average ≥ no. of dir. objects longer than average (often =)
  - other statistical forms are possible, too

Phase 3 and the right type of format for the data

- For nearly all cases, it is best to store the data in the so-called case-by-variable format as defined by the following rules
  - each data point (i.e., measurement of the dependent variable) is listed in a row on its own
  - every variable (dependent or independent) or every other characteristic of a data point is recorded in a column on its own
  - the first row contains the names of all variables
  - missing data are marked as NA – do not use empty cells
  - do not use numbers for categorical variables
- additional rules (less obligatory, but still useful)
  - the first column lists the names of all data points (either just a number or a real name)
  - the variable names in the first row are all in caps
  - the non-numeric data in all other rows are all in small letters

Phase 3 and the wrong type of format for the data

Sentence	Subj	Obj
The younger bachelors ate the nice little cat	3	4
He was locking the door	1	2
The quick brown fox hit the lazy dog	4	3

CASE	SENT_NO	SENTENCE	RELATION	LENGTH
1	1	The younger bachelors ate the nice little cat	subj	3
2	1	The younger bachelors ate the nice little cat	obj	4
3	2	He was locking the door	subj	1
4	2	He was locking the door	obj	2
5	3	The quick brown fox hit the lazy dog	subj	4
6	3	The quick brown fox hit the lazy dog	obj	3

Phase 4: questions to help you choose the right statistical test 1

- What kind of study is being conducted?
  - exploratory
  - hypothesis-testing
- How many and what kinds of variables are involved?
  - 1 dependent variable, → goodness-of-fit tests
  - 1 independent and 1 dependent variable → monofactorial tests for independence
  - 2+ independent variables and 1 dependent variable → multifactorial tests/analyses
  - 2+ dependent variables → multivariate analyses
- are data points related such that you can associate them with each other in a meaningful principled way?
  - no → tests for independent samples
  - yes → tests for dependent samples
  - the latter are usually more powerful

Phase 4: questions to help you choose the right statistical test 2

- What is the statistic of the dependent variable in the statistical hypothesis? (cf. above)
  - counts/frequencies, → often chi-squared tests
  - distributions → often Kolmogorov-Smirnov test
  - averages/means, → often *t*-tests
  - dispersions → often *F*-tests
  - correlations, → often *r* or *τ*
- What does the distribution of the data look like?
  - normal, → (often) parametric test
  - non-normal, → non-parametric or exact test
- How big are the samples to be collected?
  - <30, → often a risk to normality assumptions
  - ≥30, → often supporting normality assumptions

Phase 4: rejection and falsification of statistical hypotheses

- The logic of statistical testing is that of hypothesis falsification
- one does not prove that one's own H1 is correct
- one 'proves' that the H0 is wrong, which means one's H1 is right
- steps
  - 1: one defines a significance level $p_{critical}$, a threshold quantifying how quickly one will reject H0 / accept H1
  - 2: one computes the effect *e* observed in one's data (using the statistic from the statistical hypothesis)
  - 3: one computes the probability of error *p* how likely it is to find *e* if H0 is correct
  - 4: decision
    - if $p<p_{critical}$, one rejects H0 and accepts H1
    - if $p\geq p_{critical}$, one accepts / sticks to H0 and cannot accept H1
- this is not as weird as it sounds …

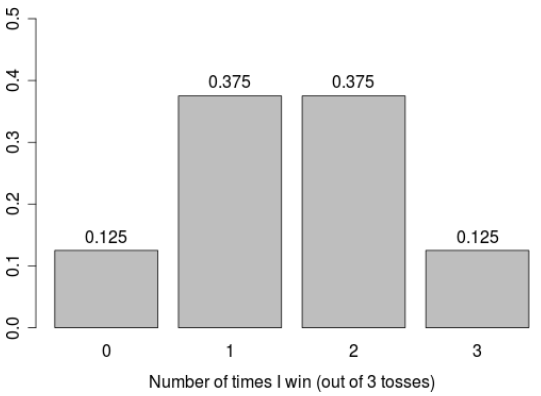
Phase 4: rejection and falsification of statistical hypotheses

- You and I play a game, tossing a coin 100 times
  - heads: $1 for me; tails: $1 for you
- your text hypotheses
  - H0: both players are honest
  - H1: STG is not honest
- your statistical hypotheses
  - H0: $p_{heads}=p_{tails}=0.5$
  - H1: $p_{heads}>0.5$ and therefore $p_{tails}<0.5$
- now, how often do you have to lose before you begin to accuse me of cheating a.k.a. accepting H1?
  - when you lose 51 times?
  - when you lose 55 times?
  - when you lose 60 times?
- what are you doing? You're looking at an effect *e* (your losses) and are determining when *e* becomes too unlikely to still believe in H0 …

Phase 4: computing *p*-values (a small example)

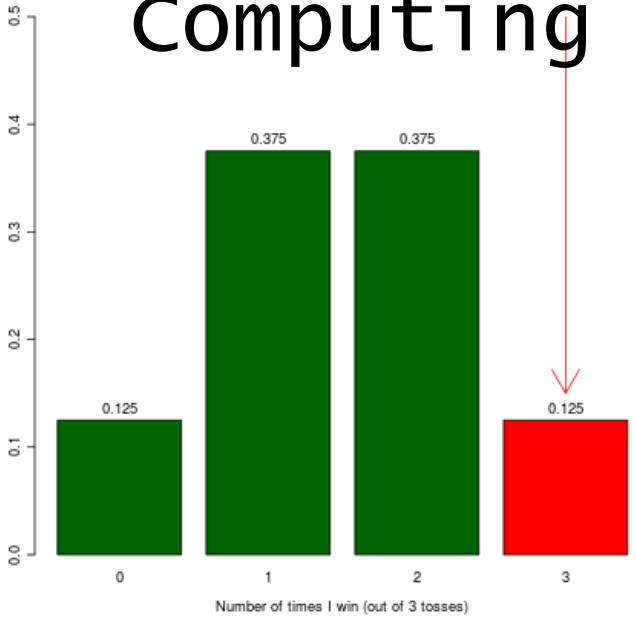
- Imagine we only tossed the coin 3 times, in which case one can just write up the whole result space

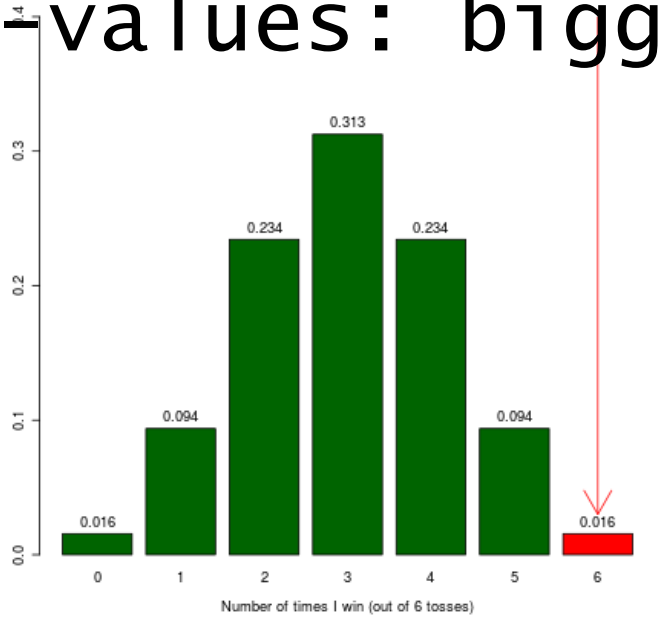
Toss 1	Toss 2	Toss 3	# heads	# tails	*p* (result)
heads	heads	heads	3	0	0.125
heads	heads	tails	2	1	0.125
heads	tails	heads	2	1	0.125
heads	tails	tails	1	2	0.125
tails	heads	heads	2	1	0.125
tails	heads	tails	1	2	0.125
tails	tails	heads	1	2	0.125
tails	tails	tails	0	3	0.125

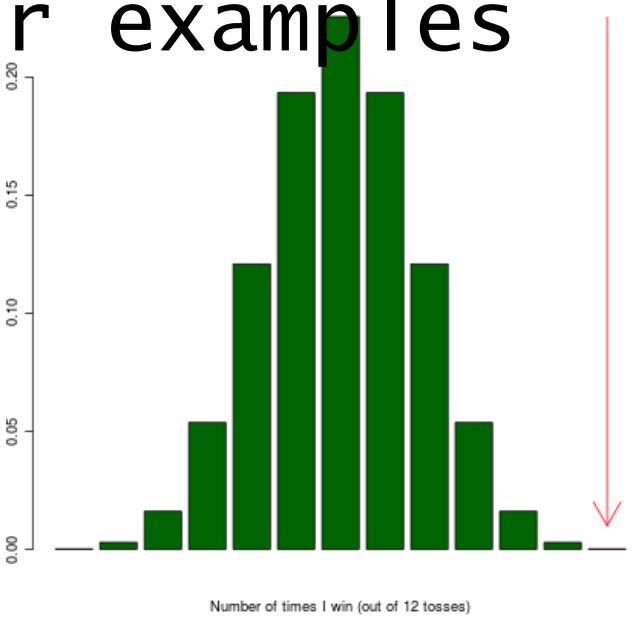


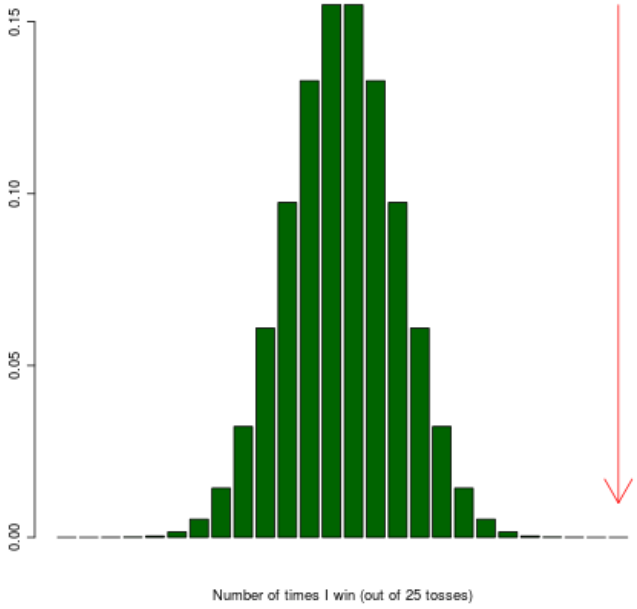
- if you lose 3 times, this is the falsificatory logic
  - 1: significance level $p_{critical}$=0.05 (5%)
  - 2: effect *e*: you won 1.5 times less than expected
  - 3: probability of error *p*=0.125 (12.5%)
  - 4: decision
    - $p \geq p_{critical}$, you accept / stick to H0 and cannot accept H1
- back to more coin tosses …

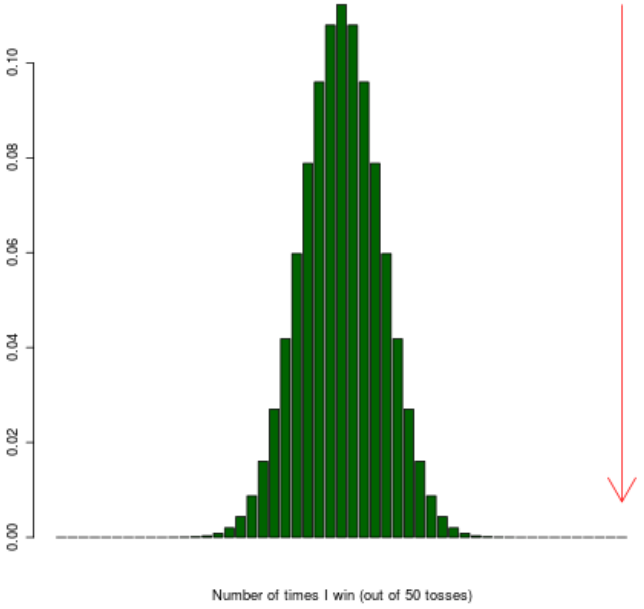
Computing *p*-values: bigger examples

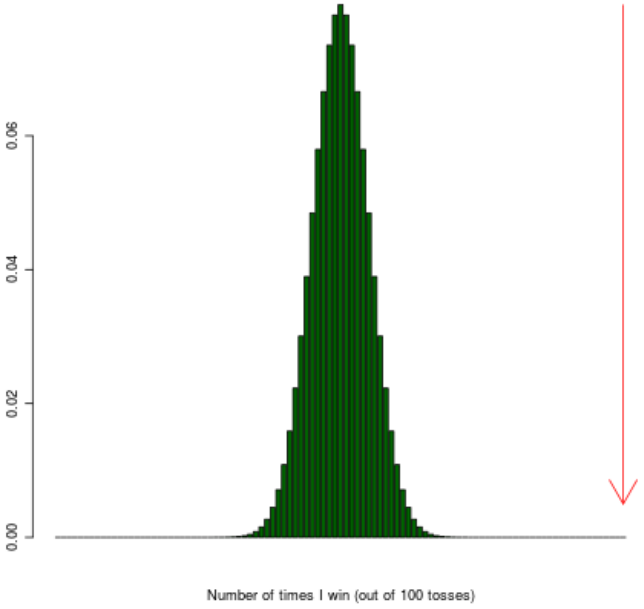












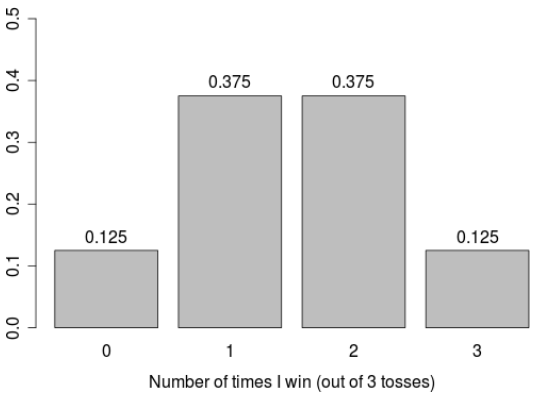
Phase 4: rejection and falsification of statistical hypotheses

- You and I play a game, tossing a coin 100 times
  - heads: $1 for me; tails: $1 for you
- an independent observer's text hypotheses
  - H0: both players are honest
  - H1: at least one player is not honest
- an independent observer's statistical hypotheses
  - H0: $p_{heads}=p_{tails}=0.5$
  - H1: $p_{heads}>0.5$ or $p_{tails}>0.5$
- now, how often does one of us have to lose before the independent observer begins to accuse the other one of cheating a.k.a. accepting H1?
  - when one of us loses 51 times?
  - when one of us loses 55 times?
  - when one of us loses 60 times?
- what is the independent observer doing? He's looking at an effect *e* (one's losses) and is determining when *e* becomes too unlikely to still believe in H0 …

Phase 4: computing *p*-values (a small example)

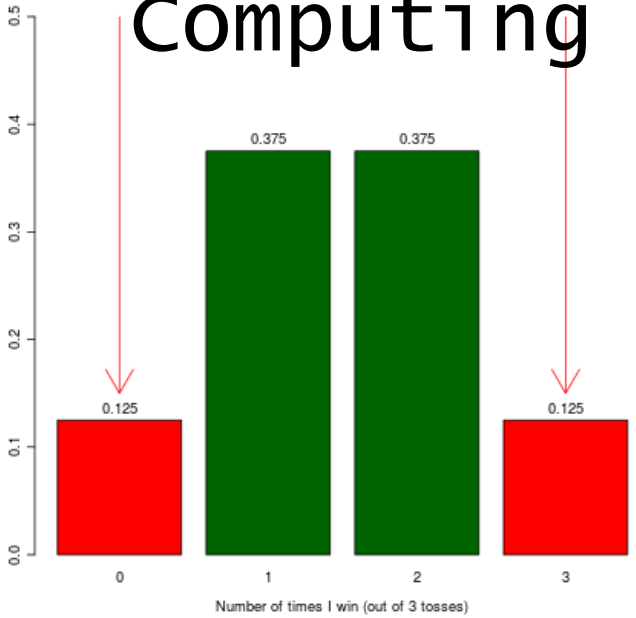
- Imagine we only tossed the coin 3 times, in which case one can just write up the whole result space

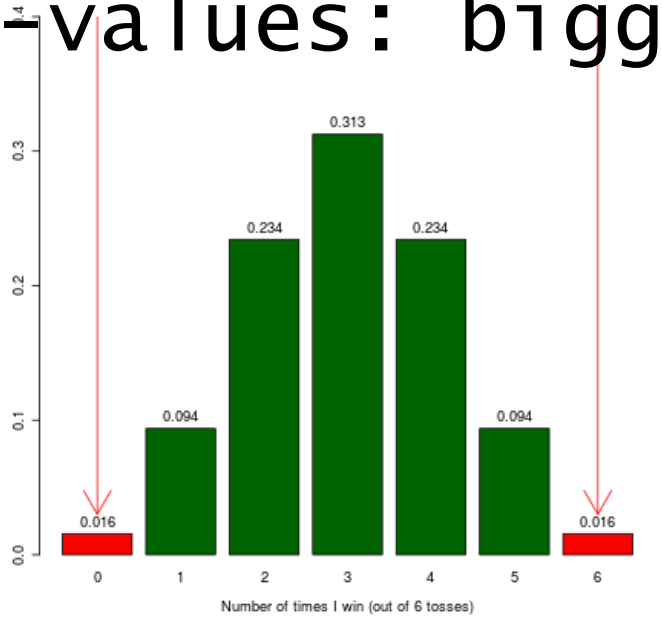
Toss 1	Toss 2	Toss 3	# heads	# tails	*p* (result)
heads	heads	heads	3	0	0.125
heads	heads	tails	2	1	0.125
heads	tails	heads	2	1	0.125
heads	tails	tails	1	2	0.125
tails	heads	heads	2	1	0.125
tails	heads	tails	1	2	0.125
tails	tails	heads	1	2	0.125
tails	tails	tails	0	3	0.125

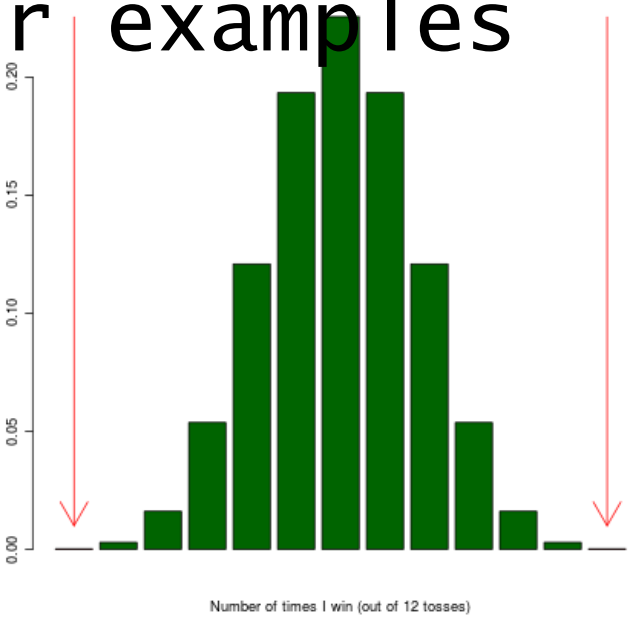


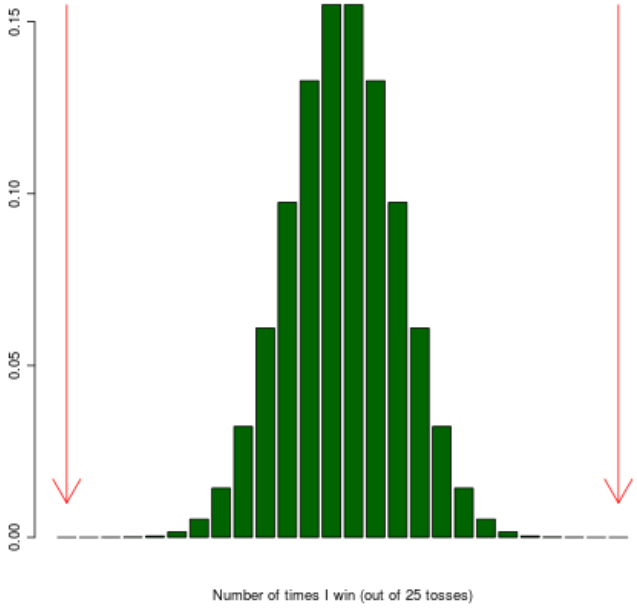
- if one of us loses 3 times, this is the falsificatory logic
  - 1: significance level $p_{critical}$=0.05 (5%)
  - 2: effect *e*: one of us won 1.5 times less than expected
  - 3: probability of error *p*=0.25 (25%)
  - 4: decision
    - $p \geq p_{critical}$, the observer sticks to H0 and cannot accept H1
- back to more coin tosses …

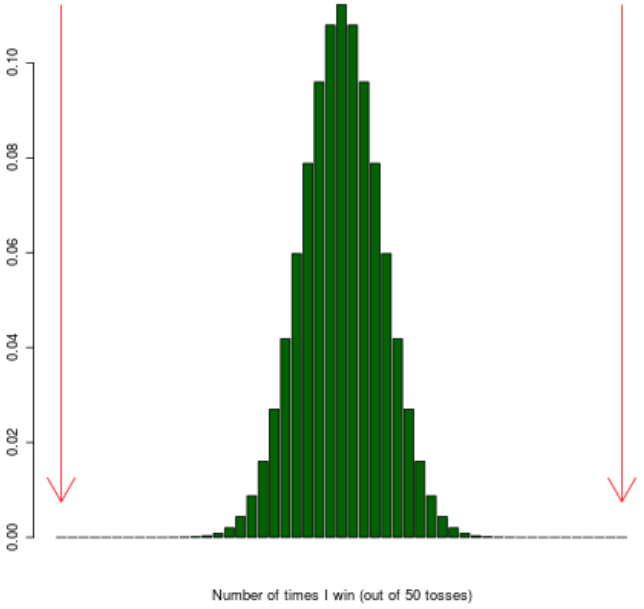
Computing *p*-values: bigger examples

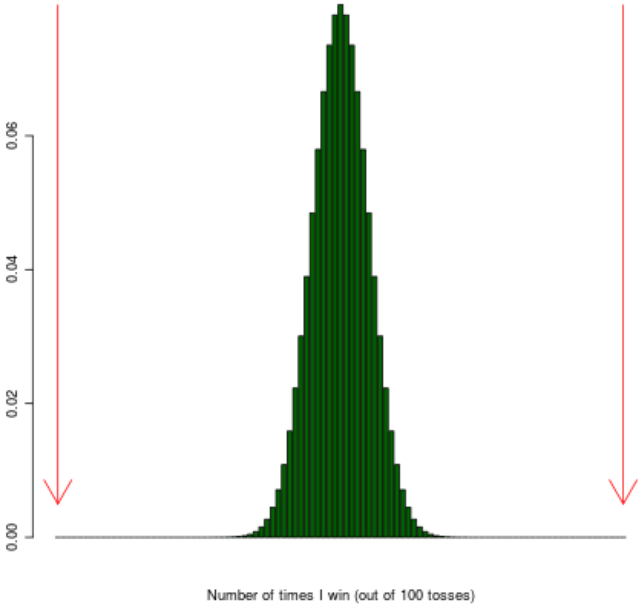












(Normal) Distributions, sample sizes, and directional hypotheses

- There are two lessons to be learned
  - on distributions and parametric testing
    - in this case of binomial trials, with increasing sample sizes, we obtain a bell-shaped normal distribution
    - thus, if the sample sizes are large enough and the distribution looks like one we can describe easily, then …
    - we can use a parametric/asymptotic test – but only then!
  - on alternative hypotheses: there are
    - non-directional/two-tailed alternative hypotheses, which postulate an effect, a difference, a correlation, *but not* its direction (in the above examples, the independent observer)
    - directional/one-tailed alternative hypotheses, which postulate an effect, a difference, a correlation *and* its direction (in the above examples, you)
    - prior knowledge is rewarded: the latter are easier to accept

Phase 4: evaluation and decision

- Significance
  - again, the *p*-value indicates how likely the observed result is, given H_0 – nothing else!

Phase 4: evaluation and decision

Levon, Erez. 2010. Organizing and processing your data: the nuts and bolts of quantitative analyses. In Lia Litosseliti (ed.). /Research methods in linguistics/. London & New York: Continuum.

Recall that the standard p-value required in the humanities and social sciences is 0.05. When we look at the relevant requirement for this p-value, we see that we need to have a chi-square statistic that is at least 5.991. With our chi-square value of 66.9, we go above and beyond this requirement, and thus can claim statistically significant findings.[3]

What does this statistical significance mean? It means that there is at least a 95% chance that the null hypothesis is *incorrect*. That indicates that we have *quantitative* support for our experimental hypothesis that educational and functional level in English affects speakers' use of null non-specific indefinite articles. If we were writing up this result in an essay or presenting it in an aca-

Phase 4: evaluation and decision

- Significance
  - again, the *p*-value indicates how likely the observed result is, given H_0 – nothing else!
  - often, levels of significance are distinguished
    - $0.1 > p \geq 0.05$ → marginally significant ms
    - $0.05 > p \geq 0.01$ → significant *
    - $0.01 > p \geq 0.001$ → very significant **
    - $p < 0.001$ → highly significant ***
- effect sizes are correlated with *p*-values, but not deterministically so - often
  - strong effects will be significant, and
  - weak effects will be insignificant, but
  - given large sample sizes, even very weak effects can be significant

	♂	♀	totals
spatl	20	15	35
socl	15	20	35
totals	35	35	70

ns

	♂	♀	totals
spatl	200	150	350
socl	150	200	350
totals	350	350	700

***