Contents

Chapter 2

Univariate Probability

This chapter briefly introduces the fundamentals of univariate probability theory, density

trials. For a frequentist, to say that $\mathbf{P}(\text{Heads}) = \frac{1}{2}$

2.4 Conditional Probability, Bayes' rule, and Independence

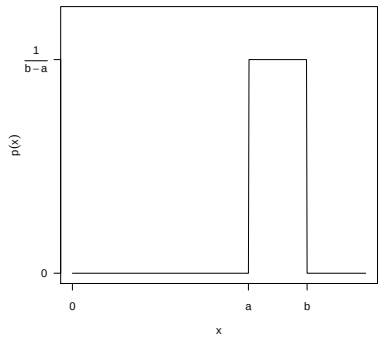
The

2.4.1 Bayes' rule

Bayes' rule

We have been given the value of the two terms in the numerator, but let us leave the

one can apply either to avoid this computation or to drastically simplify it; you will see several examples of these tricks later in the book.



(a) Probability density function

1000 B.C.E. to 500 B.C.E. (Beyer, 1986). With only this information, a crude estimate of the dis

2.8 Normalized and unnormalized probability distributions

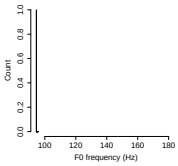
$$F(SOV) = \gamma_1\gamma_2\gamma_3$$

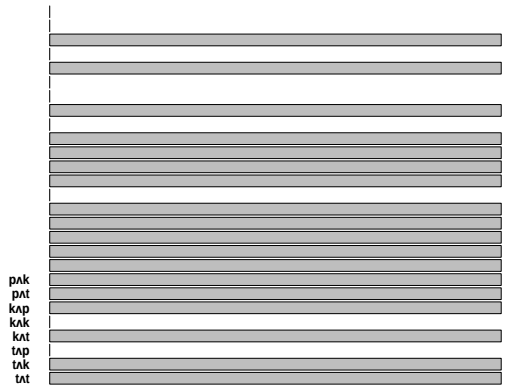
Variance of Bernoulli and uniform distributions

The variance of a Bernoulli-distributed random variable needs to be calculated explicitly, by using the definition in Equation (2.19) and summing over the possible outcomes as in

The normal distribution doesn’t have a closed-form cumulative density 5-0.191961(i03(uu)0.216467(o)-0

	⟨SB, DO⟩	**⟨SB, IO⟩**	**⟨DO, SB⟩**	**⟨DO, IO⟩**	**⟨IO, SB⟩**	**⟨IO, DO⟩**	Total
Count	478	59	1	3	20	9	570


Count
0.0
0.2
0.4
0.6
0.8
1.0
100
120
140
160
180
F0 frequency (Hz)


pʌk
pʌt
kʌp
kʌk
kʌt
tʌp
tʌk
tʌt

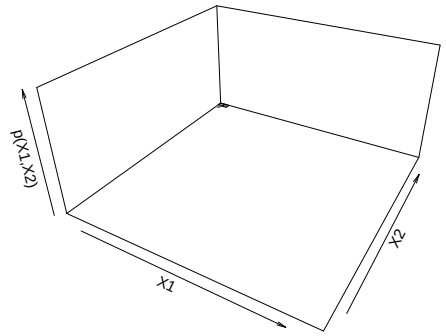
$$\mathbf{P}(\mathbf{y}|\hat{\mathbf{P}}) = \prod^{\mathbf{n}}$$

3.3 Linearity of expectation, covariance, correlation, and variance sums of random variables

Since the covariance between conditionally independent random variables is zero, it follows that the variance of the sum of pairwise independent random variables is the sum of their variances.

3.4 The binomial distribution

a sequence of r random variables $X_1, \ldots, X_r$ whose joint distribution is characterized by r parameters: a size parameter n denoting the number of trials, and $r - 1$ parameters $\pi_1, \ldots, \pi_{r-1}$, where π_i


p(x1,x2)
x1
x2

i i

E E U

a

a U

$$P(X = k; r, p) = \binom{a}{b} (1-p)^c p^d, k \in \{r, r+1, \cdots\}$$

for some choice of a, b, c, d. Complete the specification of the distribution (i.e., say what a, b, c, d are) and justify it.

Exercise 3.8: Linearity of expectation

You put two coins in a pouch; one coin is weighted such that it lands heads [5]

Chapter 4

Parameter Estimation

Thus far we have concerned ourselves primarily with *probability theory*: what events may occur with what probabilities, given a model family and choices for the parameters. This is useful only in the case where we know the precise model family and parameter values for the

for passivization will in fact be realized as a passive.

4.2.1 Consistency

An estimator is **consistent** if the estimate $\hat{\theta}$ it constructs is guaranteed to converge to the true parameter value θ as the quantity of data to which it is applied increases. Figure 4.1 demonstrates that Estimator 1 in our example is consistent: as the sample size increases, the

only the first **n/**

(Tool, 1949, cited in *Language Log* by Benjamin Zimmer, 18 October 2007)

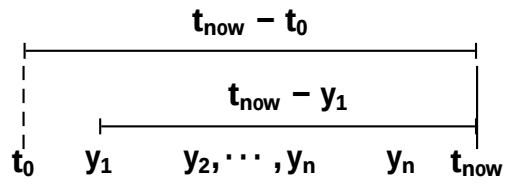
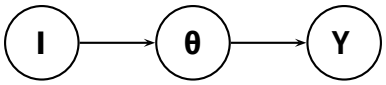
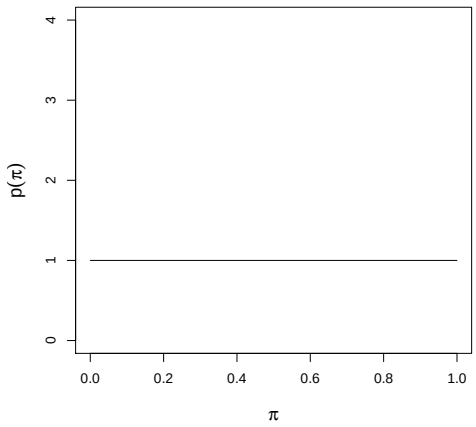


Figure 4.3: The bias of the MLE for uniform distributions

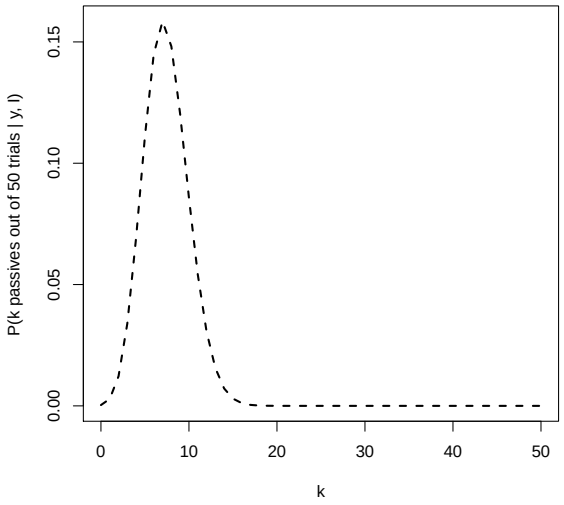

I
θ
Y

$$P(\theta|y, I) =$$


p(π)
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
π

given transitive sentence will be in the passive voice. For Bayesian statistics, we must first

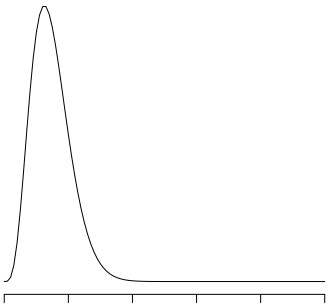
we covered a moment ago) and the *mean*. For our example, the posterior mode is $\frac{4}{}$


P(k passives out of 50 trials | y, I)
0.00
0.05
0.10
0.15
0
10
20
30
40
50
k

```
> plot(density(res[[1]][, 1]), xlab = expression(pi), ylab = expression(paste("p(",
+     pi, ")")))
```

simply expresses that observations $\mathbf{y}$ are drawn from a normal distribution parameterized by $\boldsymbol{\mu}$

dous modeling flexibility. The only real limits are conceptu



testing in particular works just like any other type of Bayes

In our second hypothesis

$$P(y|H_2) = \sum_i P(y|\pi_i) P(\pi_i|H_2)$$

$$= P\left(y|\pi = \frac{1}{3}\right) P\left(\pi = \frac{1}{3}|H_2\right) + P\left(y|\pi = \frac{2}{3}\right) P\left(\pi = \frac{2}{3}|H_2\right)$$

$$= \binom{6}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 \times 0.5 + \binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2 \times 0.5$$

$$= 0.21$$

thus

$$P(y) = P(y|H_1) P(H_1) + P(y|H_2) P(H_2)$$

$$= 0.23 \times 0.5 + 0.21 \times 0.5$$

We use the critical trick of recognizing this integral as a beta function (Section 4.4.2), which gives us:

$$= \binom{6}{4} \mathbf{B}(5,$$

5.2.3 Example: Learning contextual contingencies in sequences

Here's an example, where we will explain the **standard error of the mean**. Suppose

1. The null hypothesis is true, but we reject it (probability

Quantifying association: odds ratios

In Section 3.3 we already saw one method of quantifying the strength of association between two binary categorical variables:

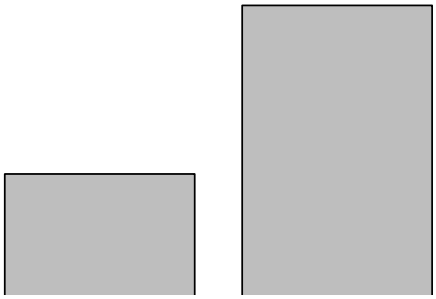
Fisher’s exact test

Fisher’s exact test applies to 2 ×

3. The linguist writes up her research results and sends them to a prestigious journal.

(c) B B B B A A A A B B B B B A A A A B B B B

The file spillover_word_rts



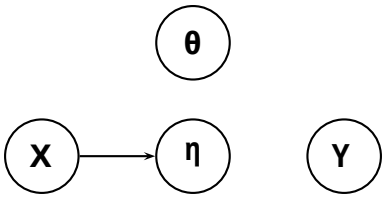
Ngarrka- ngku ka wawirri panti- rni. (Hale, 1983)
man **erg aux** kangaroo spear **nonpast**

“The man is spearing the kangaroo”.

In some dialects of Warlpiri, however, using the ergative case is not obligatory. Note

Chapter 6

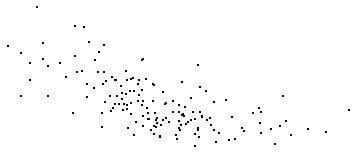
Generalized Linear Models


θ
X
η
Y

y

×

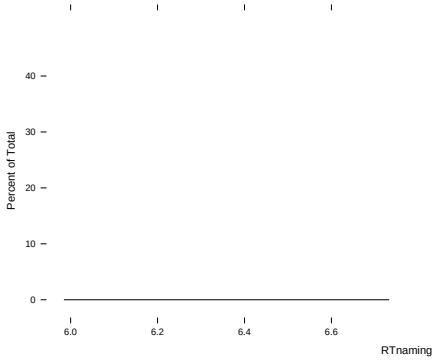
$$y_j - \hat{y}_j$$



Observed data

respectively to the boxes $\mathbf{M_A - M_0}$ and Unexplained. Thus, using the $\mathbf{F}$ statistic for hy-

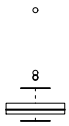
mcmrcl70n factor could have arbitrarily di erent c


40
30
20
10
0
Percent of Total
6.0
6.2
6.4
6.6
RTnaming

Frication	Age	X_1	X_2	X_3	X_4	X_5	X_6	X_7
burst	old	0	0	0	0	0	0	0
frication	old	1	0	0	0	0	0	0
long	old	0	1	0	0	0	0	0
short	old	0	0	1	0	0	0	0
burst	young	0	0	0	1	0	0	0
frication	young	1	0	0	1	1	0	0
long	young	0	1	0	1	0	1	0

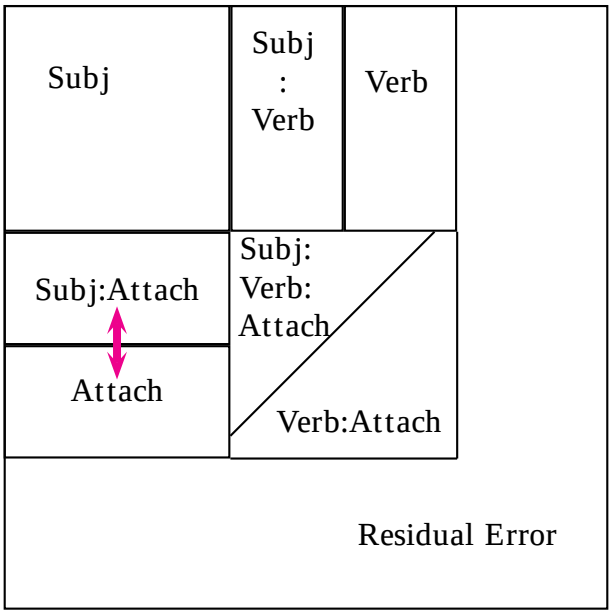
an extremely rich topic, and we take them up in Chapter 8 in full detail. There is also, however, a body of analytic techniques which uses the partit

precise reason for this. Suppose that we were to test for the presence of an interaction


8

```
+     result
+   }
> get.z.score <- function(response,conds.list) {
+     means <- tapply(response,conds.list,mean)
```

		Subject					
Verb	Attachment	1	2	3	4	5	...


Subj
Subj
:
Verb
Verb
Subj:Attach
Attach
Subj:
Verb:
Attach
Verb:Attach
Residual Error

Error: subj:verb:attachment

```
452.2940 482.0567
> with(sp.1.subj,tapply(rt,list(verb,attachment),mean))
          high      low
IC    430.4316 474.1565
```

So sentences with nonpronominal recipients are realized ro

is approximately distributed as a χ^2_k random variable, where k

6.9 Log-linear and multinomial logit models

Onset	Freq	P_{M_1}	P_{M_2}	P_{M_3A}	P_{M_3B}	Onset	Freq	P_{M_1}	P_{M_2}	P_{M_3A}	P_{M_3B}

and idiosyncratically to the probability of other words wit


λ_2
λ_1

[sr], for example, would now have the paired-segment feature sr

covered in Section XXX. In this model, there is a collection of feature functions $\mathbf{f_j}$ each of which maph

This is a new expression of the same model, but with fewer para

is McCullagh and Nelder (1989). For GLMs on categorical data, Agresti (2002) and the more introductory Agresti (2007) are highly recommended. For more information specific to

- Word frequency
- Speaker sex
- Speech rate

```
> lexdec.lm <- lm(RT ~ Frequency, lexdec)
> summary(lexdec.lm)[[4]]
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  6.58877844 0.022295932 295.514824 0.000000e+00
Frequency   -0.04287181 0.004532505  -9.458744 1.026564e-20
> summary(lexdec.lm)[[4]][2,4]
[1] 1.026564e-20
```

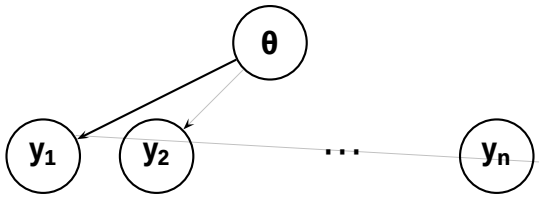
Exercise 6.5: Decomposition of variance

$\boldsymbol{\beta} = 0$ against an alternative-hypothesis model $\mathbf{M_A}$ with unconstrained $\boldsymbol{\beta}$ b $_{\mathrm{A}}$

	(Pro)noun	9192
	Verb	904
	Coordinator	1199
	Number	237
	(Pre-)Determiner	3427
	Adverb	1846
	Preposition or Complementizer	2418
1	*wh*-word	

Chapter 8

Hierarchical Models


θ
y1
y2
...
yn

1. Construct point estimates of the parameters of interest, Σ_b and θ, using the principle

	lower bound	upper bound	posterior mode
μ	613.6	644.7	618.9
σ_b	35.4	60.4	43.7
σ_y	13.5	20.9	19.1

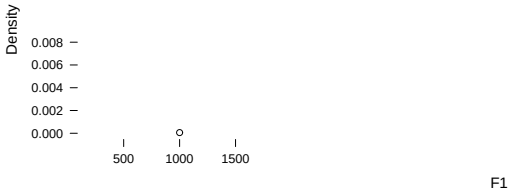
parameter (as indicated by the

Fully Bayesian analysis

We can try a similar analysis using fully Bayesian techniques rather than the point estimate. We'll present a slightly simpler model in which the speaker-

study of language. We move from generalized linear models (G

To illustrate the approach, we construct a model with the length, animacy, discourse accessibility, pronominality, and definiteness of both the recipient and theme arguments as predictors, and with verb as a random e ect. We use log-transformed length predictors (see Section 6.7.4 for discussion).

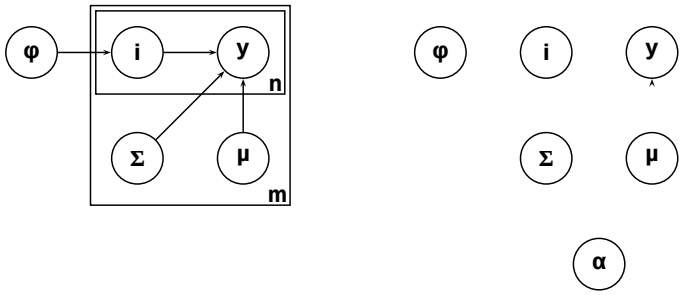

Density
0.008
0.006
0.004
0.002
0.000
500
1000
1500
F1

safely conclude that the e ects of these factors truly are equal and opposite? (**Hint:** the easiest way to construct the simpler model is to define new quanti

estimated density, and overlay

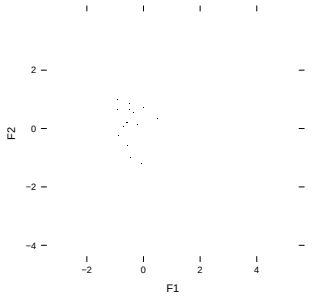
Chapter 9

Dimensionality Reduction and Latent Variable Models

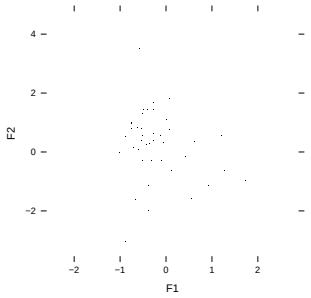


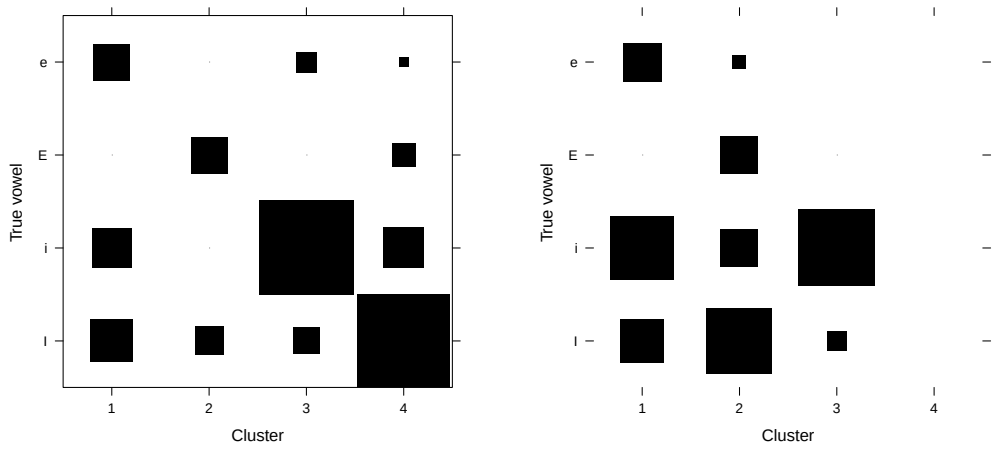
(a) Simple model

9.1.1 Inference for Gaussian mixture models


2
0
−2
−4
F2
−2
0
2
4
F1

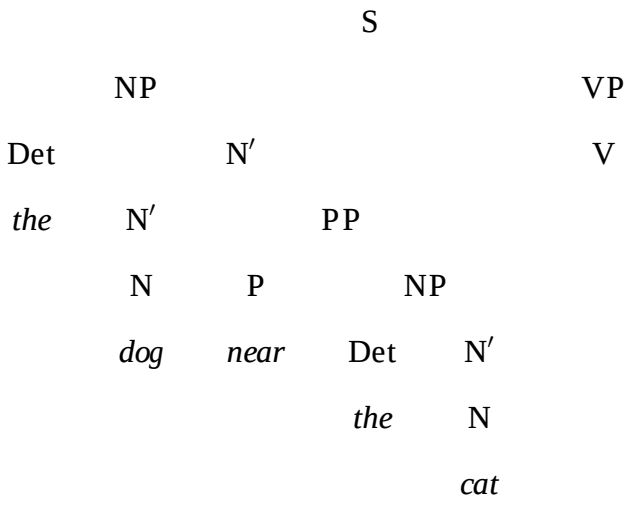
e
E
i
I
True vowel
Cluster


4
2
0
−2
F2
−2
−1
0
1
2
F1



(a) With learning of category frequencies

$$P(z_{ij} | w_{-ij}, z - ij, \sigma_\theta, \sigma_\varphi) =$$


S
NP
VP
Det
N′
V
the
N′
PP
N
P
NP
dog
near
Det
N′
the
N
cat

probabilities as follows: for each rule, write a tree-search expression corresponding using a `tgrep2`, `Tregex`, or `TIGERSearch`

Appendix A

Mathematics notation and review

This appendix gives brief coverage of the mathematical notation and concepts that you'll encounter in this book. In the space of a few pages it is of course impossible to do justice to topics such as integration and matrix algebra. Readers interested in strengthening their fundamentals in these areas are encouraged to consult XXX [calculus] and Healy (2000).

A.1 Sets ({}, □, ∩, □)

a six, with the other five outcomes all being equally likely (i.e. 10% each). If we define a discrete random variable $\mathbf{X}$ representing the outcome of a roll of this die, then the clearest way of specifying the probability mass function for $\mathbf{X}$ is by splitting up the real numbers

$$\int_a^b [f(x$$

it doesn't have the normalizing constant $\frac{1}{\sqrt{2\pi\sigma^2}}$. In order to determine the value of this

A.7 Combinatorics (

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The $\mathbf{n \times n}$ identity matrix is sometimes notated as $\mathbf{I_n}$; when the dimension is clear from context, sometimes the simpler notation $\mathbf{I}$ is used.

Transposition: For any matrix $\mathbf{X}$ of dimension $\mathbf{m \times n}$, the **transpose** of $\mathbf{X}$, or

Appendix C

but we can use the following conditional independencies, which can be read o the connec-


C

Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82(2):233–278.

Beyer, K. (1986). *The Aramaic language, its distributions and subdivisions*. Vandenhoeck & Ruprecht.

Lisker, L. and Abramson, A. S. (1967). Some e ects of context on vo

Ross, J. R. (1967). *Constraints on Variables in Syntax*