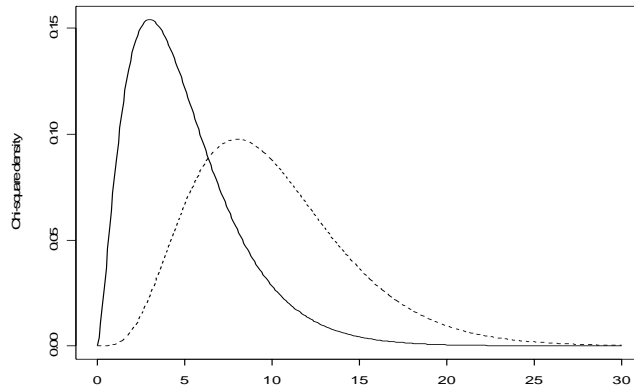


## Multivariate Methods

### Categorical Data Analysis



<http://www.isrec.isb-sib.ch/~darlene/EMBnet/>



EMBnet Course - Introduction to Statistics for Biologists 22 Jan 2009

## Variables (review)

- Statisticians call characteristics which can differ across individuals *variables*
- Types of variables:
  - *Numerical*
    - *Discrete* - possible values can differ only by fixed amounts (most commonly counting values)
    - *Continuous* - can take on any value within a range (e.g. any positive value)
  - *Categorical*
    - *Nominal* - the categories have names, but no ordering (e.g. eye color)
    - *Ordinal* - categories have an ordering (e.g. 'Always', 'Sometimes', 'Never')

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists 22 Jan 2009

## Categorical data analysis

- A categorical variable can be considered as a *classification* of observations
- Single classification
  - goodness of fit
- Multiple classifications
  - *contingency table*
  - homogeneity of proportions
  - independence

Lec 4b


EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Mendel and peas

- Mendel's experiments with peas suggested to him that seed color (as well as other traits he examined) was caused by two different 'gene alleles' (he didn't use this terminology back then!)
- Each (non-sex) cell had two alleles, and these determined seed color:

$y/y, y/g, g/y \rightarrow$  

$g/g \rightarrow$  

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Peas, cont

- Here, yellow is dominant over green
- Sex cells each carry one allele
- Also postulated that the gene pair of a new seed determined by combination of pollen and ovule, which are passed on *independently*

pollen parent		seed parent	
Y	g	Y	g
YY	Yg	gY	gg
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Did Mendel's data prove the theory?

- We know today that he was right, but how good was his experimental proof?
- The statistician R. A. Fisher claimed the data fit the theory *too well*:
  - 'the general level of agreement between Mendel's expectations and his reported results shows that *it is closer than would be expected* in the best of several thousand repetitions.... I have no doubt that Mendel was deceived by a gardening assistant, who know only too well what his principal expected from each trial made'
- How can we measure how well data fit a prediction?

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Testing for goodness of fit

- The NULL is that the data were generated according to a particular chance model
- The model should be *fully specified* (including parameter values); if parameter values are not specified, they may be estimated from the data
- The TS is the *chi-square statistic* :  
$$\chi^2 = \text{sum of } [(observed - expected)^2 / expected]$$
- The  $\chi^2$  distribution depends on a number of *degrees of freedom*

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Example

- A manager takes a random sample of 100 sick days and finds that 26 of the sick days were taken by the 20-29 age group, 37 by 30-39, 24 by 40-49, and 13 by 50 and over
- These groups make up 30%, 40%, 20%, and 10% of the labor force at the company. Test the hypothesis that age is not a factor in taking sick days ...

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Example, contd

Age	Observed	Expected	Difference	$\chi^2$
20-29	26	.3*100=30	26-30=-4	$(-4)^2/30$ =.533
30-39	37			
40-49	24			
≥ 50	13 (total=100)			

- $\chi^2 = .533 + \underline{\quad} + \underline{\quad} + \underline{\quad} \approx 2.46$
- To get the p-value in R:  
`> pchisq(2.46, 3, lower.tail=FALSE)`

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Multiple variables: rxc contingency tables

- A *contingency table* represents all combinations of variable levels for the different classifications
- r = number of rows, c = number of columns
- *Example:*
  - Hair color = Blond, Red, Brown, Black
  - Eye color = Blue, Green, Brown
- Numbers in table represent *counts* of the number of cases in each combination ('cell')
- Row and column *totals* are called *marginal counts*

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

### Hair/eye table

Eye \ Hair	Blue	Green	Brown	
Blond	$n_{11}$	$n_{12}$	$n_{13}$	$n_{1.}$
Red	$n_{21}$	$n_{22}$	$n_{23}$	$n_{2.}$
Brown	$n_{31}$	$n_{32}$	$n_{33}$	$n_{3.}$
Black	$n_{41}$	$n_{42}$	$n_{43}$	$n_{4.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	Grand Total $n_{..}$

cells

row margins

column margins

Lec 4b      EMBnet Course - Introduction to Statistics for Biologists      22 Jan 2009

### Hair/eye table for our class

Eye \ Hair	Blue	Green	Brown	
Blond				
Red				
Brown				
Black				

Lec 4b      EMBnet Course - Introduction to Statistics for Biologists      22 Jan 2009

## Special Case: 2x2 tables

- Each variable has *2 levels*
- Measures of *association*
  - Odds ratio (cross-product)  $ad/bc$
  - Relative risk  $[ a/(a+b) / (c/(c+d)) ]$

	+	-	Total
group 1	a ( $n_{11}$ )	b ( $n_{12}$ )	$n_{1.}$
group 2	c ( $n_{21}$ )	d ( $n_{22}$ )	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Chi-square Test of Independence

- Tests association between two categorical variables
  - NULL: The 2 variables (classifications) are independent
- Compare observed and expected frequencies among the cells in a contingency table
- The TS is the *chi-square statistic*:  
 $\chi^2 = \text{sum of } [(observed - expected)^2 / expected]$
- $df = (r-1)(c-1)$ 
  - So for a 2x2 table, there is *1 df*

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Chi-square independence test: intuition

- Construct bivariate table as it would look under the NULL, ie if there were *no association*
- Compare the *real table* to this hypothetical one
- *Measure how different* these are
- If there are *sufficiently large differences*, we conclude that there is a *significant relationship*
- Otherwise, we conclude that our numbers vary just due to chance

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Expected frequencies

- How do we find the *expected* frequencies?
- Under the NULL hypothesis of independence, the chance of landing in any cell should be the *product of the relevant marginal probabilities*
- ie, expected number  $n_{ij}$   
$$= N \cdot [(n_{i.}/N) \cdot (n_{.j}/N)]$$
$$= n_{i.} \cdot n_{.j} / N$$

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Are hair and eye color independent?

- Let's see...

## Chi-Square test assumptions

- Data are a *simple random sample* from some population
- Data must be *raw frequencies* (not percentages)
- Categories for each variable must be mutually exclusive (and exhaustive)
- The chi-square test is based on a large sample approximation, so the *expected numbers should not be too small* (at least 5 in most cells)

## Another Example

- Quality of sleep before elective operation...

	Bad	OK	Total
trt	2	17	19
Placebo	8	15	23
Total	10	32	42

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## A lady tasting tea

- Exact test developed for the following setup:
- A lady claims to be able to tell whether the tea or the milk is poured first
- 8 cups, 4 of which are tea first and 4 are milk first (and the lady knows this)
- Thus, the margins are known in advance
- Want to assess the chance of observing a result (table) *as or more extreme*

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Fisher's Exact Test

- Method of testing for association when some *expected values are small*
- Measures the chances we would see differences of this magnitude or larger if there were *no association*
- The test is *conditional on both margins* - both the row and column totals are considered to be *fixed*

## More about Fisher's exact test

- Fisher's exact test computes the probability, given the observed marginal frequencies, of obtaining exactly the frequencies observed and any configuration more extreme
- '*More extreme*' means any configuration with a *smaller probability of occurrence* in the same direction (one-tailed) or in both directions (two-tailed)

## Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

## Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

	+	-	
A	0		5
B			10
	8	7	15

	+	-	
A	1		5
B			10
	8	7	15

	+	-	
A	3		5
B			10
	8	7	15

	+	-	
A	4		5
B			10
	8	7	15

	+	-	
A	5		5
B			10
	8	7	15

Lec 4b

EMBnet Course - Introduction to Statistics for Biologists

22 Jan 2009

Example

	+	-	
A	2	3	5
B	6	4	10
	8	7	15

.326

	+	-	
A	3		5
B			10
	8	7	15

.392

.007

	+	-	
A	0		5
B			10
	8	7	15

	+	-	
A	4		5
B			10
	8	7	15

.163

.093

	+	-	
A	1		5
B			10
	8	7	15

	+	-	
A	5		5
B			10
	8	7	15

.019

Lec 4b      EMBnet Course - Introduction to Statistics for Biologists      22 Jan 2009

## Where do these probabilities come from??

- With both margins fixed, there is only 1 cell that can vary
- The probabilities come from the hypergeometric distribution
- This distribution gives probabilities for the number of 'successes' in a sample of size n drawn without replacement from a population of size N comprised of a known number of 'successes'
- *Chocolates...*

Lec 4b      EMBnet Course - Introduction to Statistics for Biologists      22 Jan 2009

## Fisher's exact test in R

- In R, use the command  
> `fisher.test()`
- Let's try the Fisher test on the earlier data...

## Problems with Fisher's test

- The exact test was developed for the case of fixed marginals
- In this case the probability (p-value) computed by the Fisher test is *exact* (unlike the chi-square test, which relies on approximations)
- However, this setup is unrealistic for most studies - even if we know how many samples we will get in each group, we generally cannot fix in advance *both margins*
- Other methods have also been proposed to deal with this problem

## Summary

- Multivariate data analysis can be either *descriptive* or *inferential*
- Methods depend on the *type of variables* in the data
- For categorical variables, we have looked at large sample and small sample tests of association