Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Variables : categorical and discrete data
- Goodness-of-fit
- Contingency tables
- Visualizing categorical data : mosaic plot
- Tests of independence, homogeneity
- (Cochran-) Mantel-Haenzel test
- Example : UC Berkeley Admisssions data (Simpson's paradox)

Variables

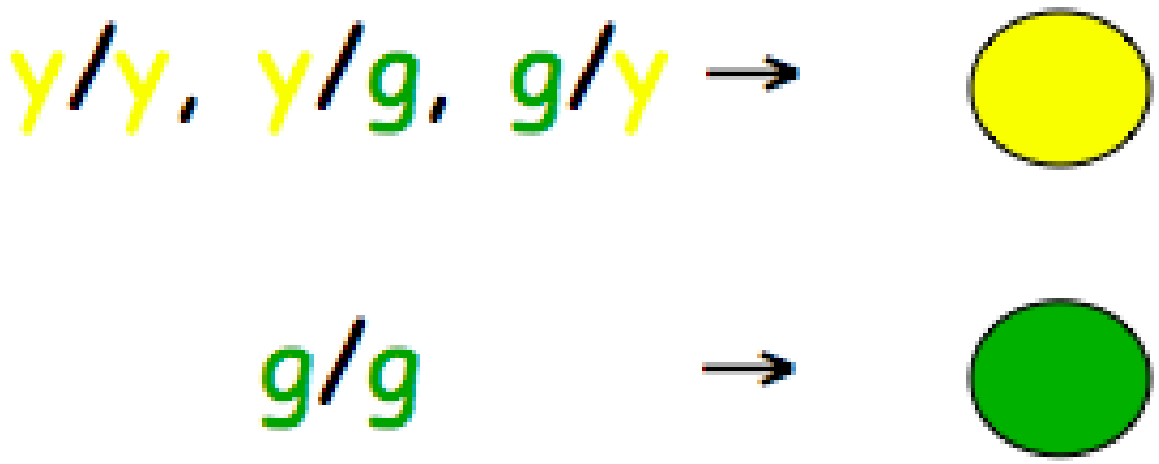
- Statisticians call characteristics which can differ across individuals *variables*
- Types of variables :
- **Numerical**
  - *Discrete* – possible values can differ only by fixed amounts (most commonly counting values)
  - *Continuous* – can take on any value within a range (e.g. any positive value)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

Categorical data analysis

- A categorical variable can be considered as a *classification* of observations
- Nonparametric, randomization-based methods
- Single classification
  - goodness-of-fit
- Multiple classifications
  - *contingency table*
  - homogeneity of proportions
  - independence
- Model-based analysis : more flexible, useful for *estimation*

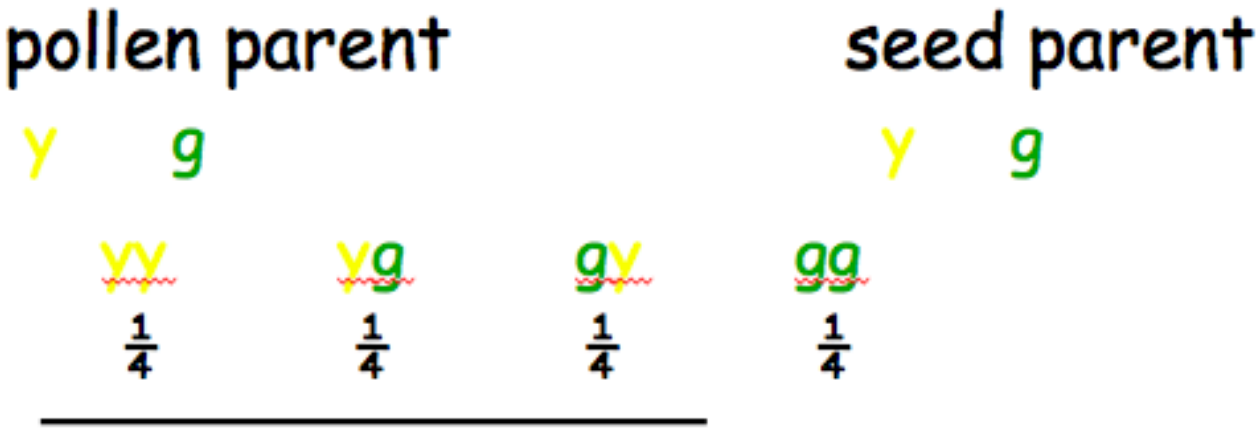
Mendel and peas

- 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 :



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*



Did Mendel's data support the theory ?

- We know today that he was right, but how good was his experimental proof ?
- → **How can we measure how well data fit a prediction ? ?**
- Want to test for *goodness-of-fit* of observed data to a theoretical distribution

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
- Test statistic : $X^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i}$
- Under the NULL, $X^2 \sim \chi^2_{k-1}$, where k is the number of categories

Example

- A manager takes a random sample of 100 sick days and finds :
  - 26 of the sick days were taken by the 20-29 age group
  - 37 by 30-39
  - 24 by 40-49
  - 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

...

Example, cont.

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

- $\chi^2 = .533 + \_\_\_\_\_ + \_\_\_\_\_ + \_\_\_\_\_ \approx 2.46$
- To get the p-value in R:

```
> pchisq(2.46,3,lower.tail=FALSE)
```

Continuous example : uniform distribution

- Consider a RV X taking values between 0 and 1, but whose density is unknown
- In addition, suppose that there is a sample of size 100 of observations from this distribution and that we would like to test whether the observations are compatible with a uniform distribution
- We could divide the interval (0,1) into 20 sub-intervals (0,0.05), (0.05,0.10), *etc.*
- If the distribution is uniform, then the probability that any particular observation is in sub-interval i is $p_i = 1/20$, $i = 1, \ldots, 20$
- So the expected number in each sub-interval is $np_i = 100/20 = 5$
- Then we can compute the statistic X^2 as above

Goodness-of-fit for continuous distributions

- We can use this same method for *any continuous distribution*
- Choice of k is somewhat arbitrary : in general, choose k and the sub-intervals such that the expected numbers are approximately equal and not too small

1. Partition the support of the distribution (finite or infinite) into k disjoint sub-intervals
2. Determine the probability p_i^0 supposing a specific distribution (H), and thus the expected values (np_i)
3. Obtain N_i, the number of observations in sub-interval i
4. Calculate X^2_{obs}
5. Sous H $X^2_{obs} \sim \chi^2_{k-1}$

Testing composite hypotheses

- We have been considering **simple** null hypotheses : that is, the distribution is *completely specified* by the hypothesis
- For example $H : p = 0.1$ for a binomial proportion is a simple hypothesis
- In this case, it is not necessary to estimate any parameters, and the statistic $X^2_{obs} \sim \chi^2_{k-1}$ under H
- However, it could happen that we are interested in null hypotheses containing *multiple possible values* for the parameter(s)
- This type of hypothesis is a *composite* hypothesis

Test statistic for composite hypotheses

- For a composite null hypothesis, we must *modify* the test statistic X^2_{obs}, since the expected number in class i is no longer completely specified by the null hypothesis H
- **Modification :** replace np_i by its MLE $n\hat{p}_i$ in calculating X^2_{obs}
- Under H, this $X^2_{obs} \sim \chi^2_{k-p-1}$, where p is the *number of estimated parameters* in calculating $\hat{p}_i$

Example : parameter estimation

- 200 (independent) indivuduals are asked how many lottery tickets they bought last week
- Results :

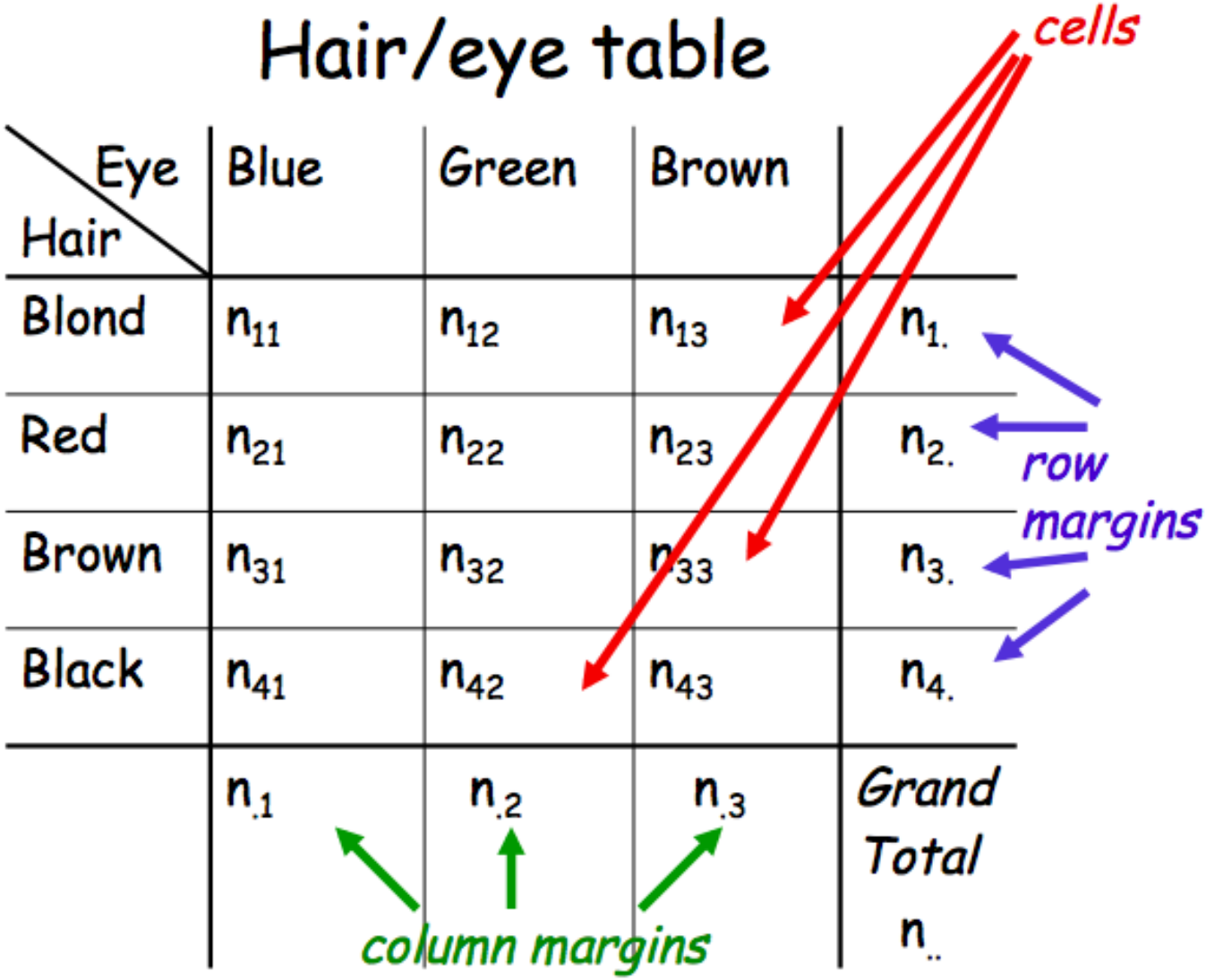
# billets	0	1	2	3	4	5	6	7	8	9	10	≥ 11
# personnes	52	60	55	18	8	3	2	1	0	0	1	0

- Test the hypothesis that these observations follow a Poisson distribution...

Contingency table

- A *contingency table* is a specific way to *simultaneously* represent 2 (or more) characteristics on a population (or sample)
- The representation consists of the *frequencies* of values for each couple of variables $(X,\ Y)$ with modalities $x_i = 1, \ldots, r$, $y_j = 1, \ldots, c$
- r = number of rows, c = number of columns
- Example :
  - Hair color = blond, red, brown, black
  - Eye color = brown, green, blue
- The values in the table represent *the number of observations for each combination* of possible values for the pair ('cell')
- The *sums* of values for a row or a column are the *marginal totals*

Example, cont.



Special case : 2×2 table

- 2 variables, each with *2 levels*
- Measures of *association*
  - Odds ratio (cross-product) : $\dfrac{ad}{bc}$
  - Relative risk : $\dfrac{a/(a+b)}{c/(c+d)}$

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

Visualizing categorical/discrete data

- Exploratory methods
  - Minimal assumptions (like non-parametric methods)
  - Show the data, not just summaries
  - Help detect patterns, trends, anomalies, suggest hypotheses
- Plots for model-based methods
  - Residual plots – departures from model, omitted terms, *etc.*
  - Effect plots – estimated probabilities of response or log odds
  - Diagnostic plots – influence, violation of assumptions
- R packages `vcd`, `vcdExtra` : visualizing categorical data

Mosaic plots

- *Mosaic plots* give a graphical representation of successive decompositions of a multi-way contingency table
- Counts are represented by *rectangles*
- At each stage of plot creation, the rectangles are split parallel to one of the two axes
- Even though there are rectangles, the important visual aspect is the *length* (we are comparing lengths, NOT areas)
- To make mosaic plot, need :
  - A contingency table containing the data
  - A preferred ordering of the variables, with the 'response' variable last

Example : classical music listening

	Education			
	High		*Low*	
	Classical Music			
Age	*Yes*	*No*	*Yes*	*No*
Old	210	190	170	730
Young	194	406	110	290

Data Order			
1	5	3	7
2	6	4	8

Example : classical music listening data entry in R

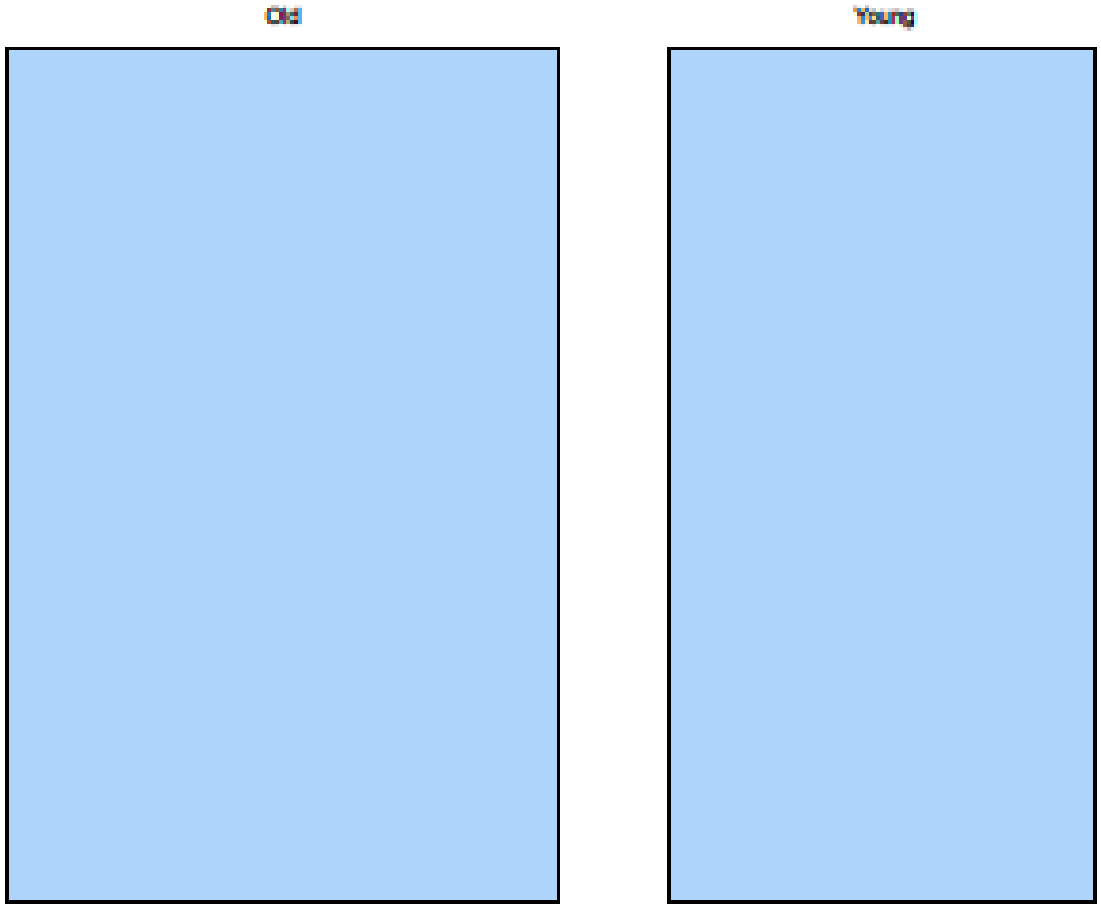
```
> music = c(210, 194, 170, 110,
            190, 406, 730, 290)

> dim(music) = c(2, 2, 2)

> dimnames(music) =
    list(Age = c("Old", "Young"),
         Education = c("High", "Low"),
         Listen = c("Yes", "No"))
```

Example, contd : looking at data in R

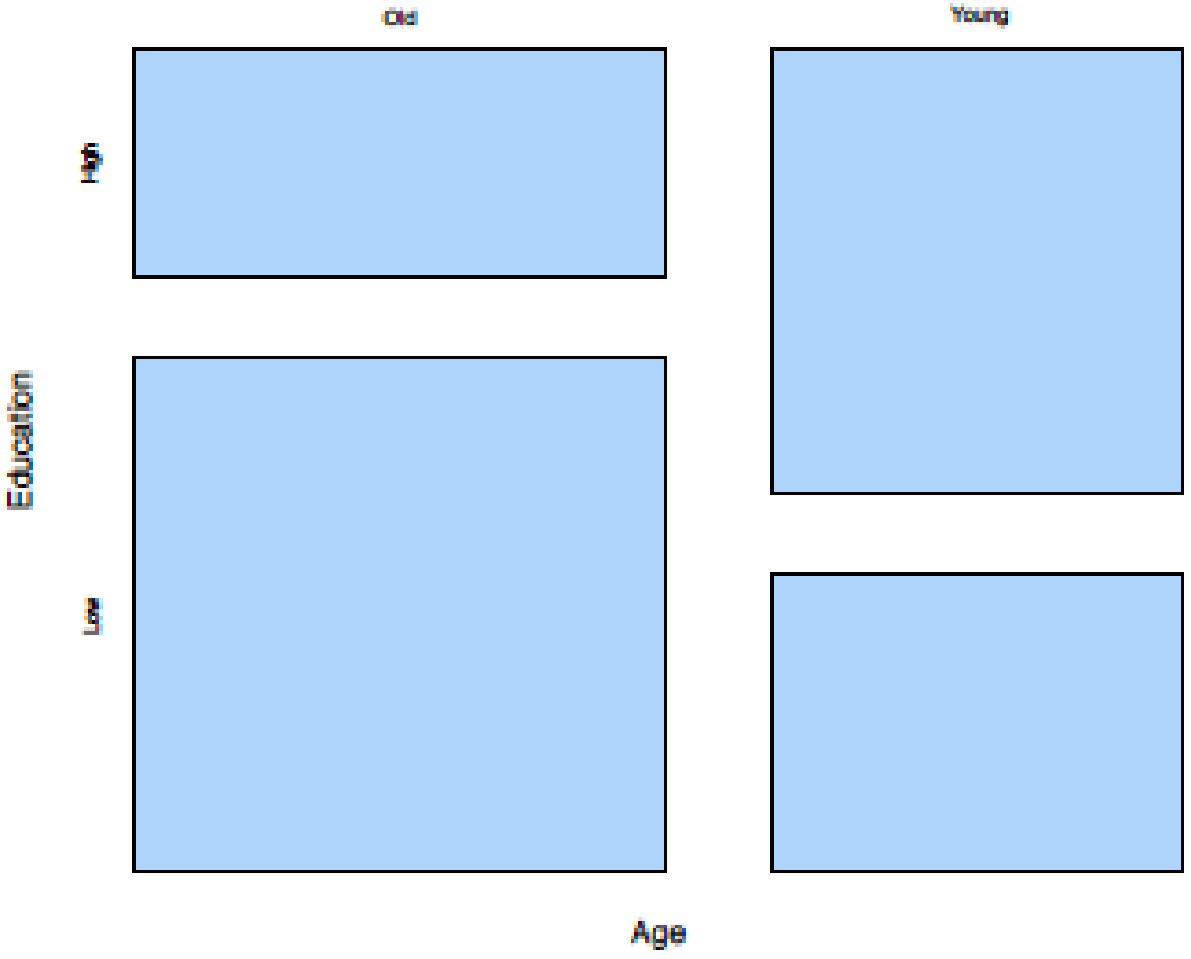
```
> music
, , Listen = Yes

         Education
Age       High Low
  Old      210 170
  Young    194 110

, , Listen = No

         Education
Age       High Low
  Old      190 730
  Young    406 290
```

Example : classical music listening

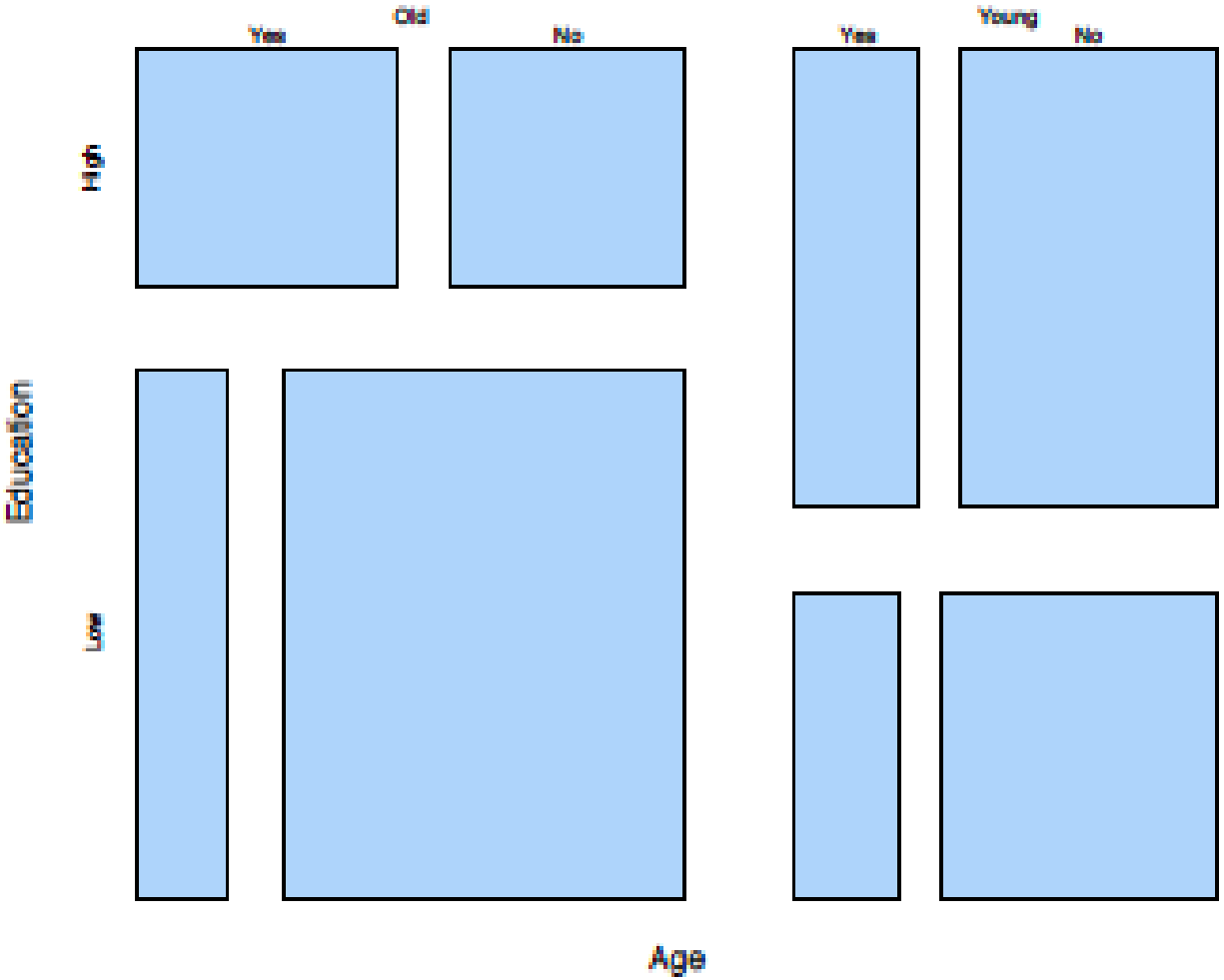
Example : classical music listening



Example : classical music listening



Example : classical music listening



BREAK

Test of independence : intuition

- Construct bivariate table as it would look under the NULL, *i.e.* if there were *no association*
- Compare the *real, observed* table to this hypothetical one
- Measure *how different* these two tables are
- If there are *sufficiently large differences*, we conclude that there is a *significant relationship* (*i.e.* a significant deviation from independence)
- Otherwise, we conclude that our numbers vary *just due to chance*

Test of independence

- 2 characteristics are *independent* if the value of one *does not influence the distribution of values of the other*
- $\Longrightarrow$ joint frequencies should be (close to) the *products of the marginal frequencies*
- Under the hypothesis that X and Y are *indeependent*, the joint probabilities are :

$$p_{ij} = P(X = x_i \text{ and } Y = y_j) = P(X = x_i) \times P(Y = y_j)$$

- Based on the contingency table, the 2 *marginal probabilities* are estimated by :

$$P(X = x_i) = \frac{\sum_{j=1}^{c} n_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} = \frac{n_{i\cdot}}{n}$$

$$P(Y = y_j) = \frac{\sum_{i=1}^{r} n_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}} = \frac{n_{\cdot j}}{n_{\cdot\cdot}} = \frac{n_{\cdot j}}{n}$$

Test of independence, cont.

- We thus obtain $\hat{p}_{ij} = n_{i.}n_{.j}/n^2$
  ($n = n_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$, the total sample size)

- $\Longrightarrow$ Under the null hypothesis of independence of row and column characters, the expected numbers in each cell (i, j) are :

$$n \times P(X = x_i) \times P(Y = y_j) = \frac{\sum_{j=1}^{c} n_{ij} \times \sum_{i=1}^{r} n_{ij}}{n} = \frac{n_{i.}\, n_{.j}}{n}$$

Test statistic

- The test statistic is still of the form :

$$X^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

$$= \frac{(n_{11} - n_{1.}n_{.1}/n)^2}{n_{1.}n_{.1}/n} + \frac{(n_{12} - n_{1.}n_{.2}/n)^2}{n_{1.}n_{.2}/n} + \cdots + \frac{(n_{rc} - n_{r.}n_{.c}/n)^2}{n_{r.}n_{.c}/n}$$

- For n 'sufficiently large' $(\forall i, j, \quad np_{ij} \geq 5)$, under H

$$X^2_{obs} \sim \chi^2_{(r-1)\times(c-1)}$$

- So for a 2×2 table, there is 1 df

Example

- The following table contains results of a survey to consider the link between *sex* and handedness *left-handed or right-handed* :

	men	women	
right-handed	2780	3281	6061
left-handed	311	300	611
	3091	3581	6672

- Test at level $\alpha = 0.05$ the hypothesis that 'sexe' et 'handedness' are independent ...

Solution

- Theoretical table :

	men	women	
right-handed	2808	3253	6061
left-handed	283	328	611
	3091	3581	6672

$2808 = 6061 \times 3091/6672$, etc.

$$X^2 = \frac{(2780 - 2808)^2}{2808} + \frac{(3281 - 3253)^2}{3253} + \frac{(311 - 283)^2}{283} + \frac{(300 - 328)^2}{328}$$

$= 5.68$

- Under H, $X^2 \sim \chi^2_1$; $\chi^2_{1,0.95} = 3.84 < 5.68$.
- Thus, we *REJECT* the null hypothesis : the data indicate *a significant deviation* from the hypothesis of independence

Test of homogeneity

- The **test of homogeneity** consists of verifying that J samples (groups) *come from the same population/distribution*
- That is, the distribution of the variable of interest is *the same* in all the populations
- This test is a *comparison of the distribution* of a qualitative variable in multiple samples
- Consider a factor A that can take I different values in the population
- The probability for each different value of A is $p_{i.}$, $i = 1, \ldots, I$
- Let J samples C_j of sizes $n_{.j}$, respectively, be taken from the population

Test of homogeneity, cont.

- The observed frequencies of the values A_i for factor A in sample C_j are noted n_{ij}
- The null hypothesis H states that the distributions are *the same*
- $\Longrightarrow$ the observed differences between all the samples are due to *sampling variability* (random variation)

Modality of factor	C_1	C_2	$\cdots$	C_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$					
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$n_{.2}$		$n_{.J}$	$n_{..} = n$

Test statistic

- Under H, *the probabilities* for the different factor modalities *have equal marginal distributions*
- Probability for modality i for factor A :

$$p_{i.} = n_{i.}/n$$

- Expected number for modality i in sample j :

$$E_{ij} = n_{.j}\, p_{i.} = \frac{n_{i.}\, n_{.j}}{n}$$

- Test statistic (homogeneity) is still of the form :

$$X^2_{obs} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

- Note : c'est *the same formula* as for the test of independence
- Under H, $X^2 \sim \chi^2_{(I-1)(J-1)}$

Small sample size : Fisher's exact test

- The χ^2 distribution for the X^2 statistic applies for 'sufficiently large' sample sizes (expected value ≥ 5 in each cell)
- *Fisher's exact test* is a method of testing for independence 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*

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* (*i.e.*, the p-value)

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 table 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

Example

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

	+	-	
A	3		5
B			10
	8	7	15

	+	-	
A	0		5
B			10
	8	7	15

	+	-	
A	4		5
B			10
	8	7	15

	+	-	
A	1		5
B			10
	8	7	15

	+	-	
A	5		5
B			10
	8	7	15

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

Cochran-Mantel-Haenzel test

- The *Cochran-Mantel-Haenszel* test is a technique that can be used to test for and/or generate an estimate of an association between an exposure and an outcome *after adjusting for or taking into account confounding*
- Used with a dichotomous outcome variable and a dichotomous risk factor, stratified by levels of the confounding factor (multiple 2×2 tables)
- Estimated odds ratio (or relative risk) is a weighted average across strata (subgroups/levels of the confounder)
- Important assumption : *no 3-way interaction*
- More on this next week

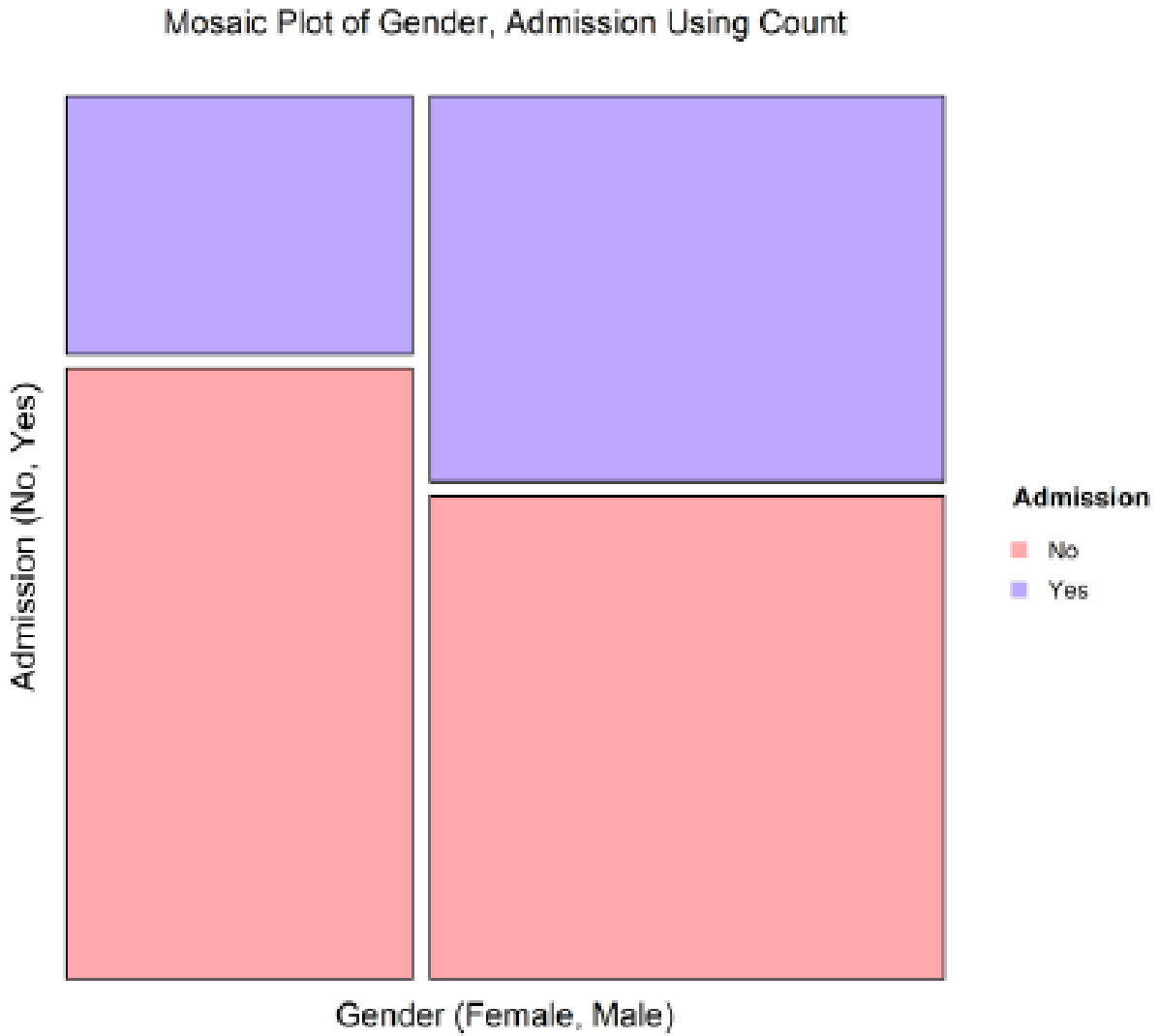
Example : Discrimination in admissions ?

- In the 1980s, a court case brought against the University of California at Berkeley by women seeking admission to graduate programs
- The women claimed that the proportion of women admitted to Berkeley was much lower than that for men, and that this was the result of discimination
- The data :

Gender	*Admitted*	*Rejected*	*%Admitted*
Male	1198	1493	44.5
Female	557	1278	30.4

- We see that a larger proportion of males is being admitted

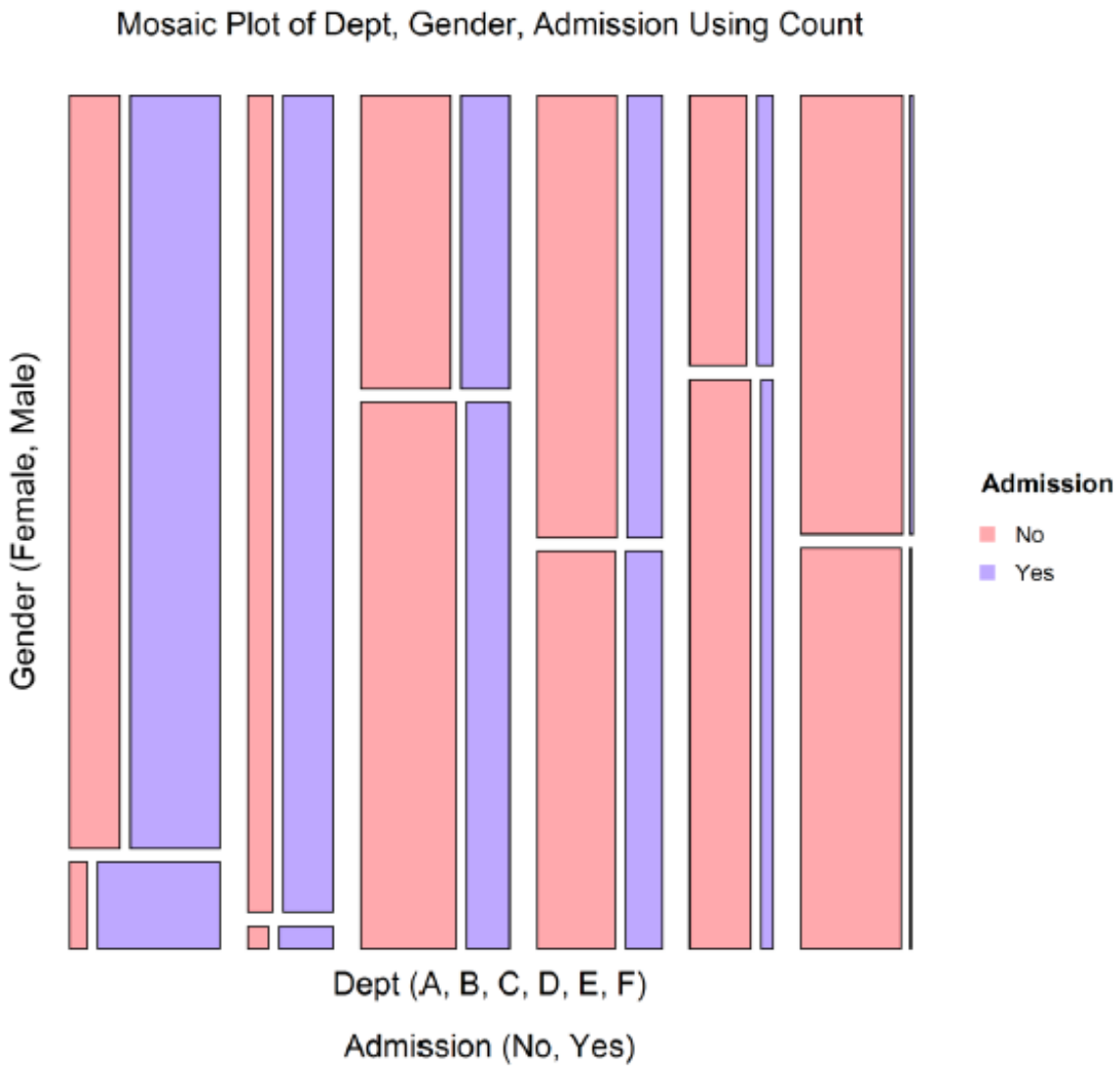
Example : mosaic plot



Example : mosaic plot

- The *widths* of the boxes are proportional to the percentage of females and males, respectively
- Here, 41% of applicants were female and 59% male
- The *heights* of the boxes are proportional to percent admitted
- In fact, 45% of the male applicants were admitted, while only 30% of the female applicants were admitted
- Boxes for those admitted are colored blue while those not admitted are colored pink
- It is easy to see that females' blue box on the left is much shorter than the males' blue box on the right
- This *seems* to show a large gender-bias in admission
- However, this inference does not take *department* into account

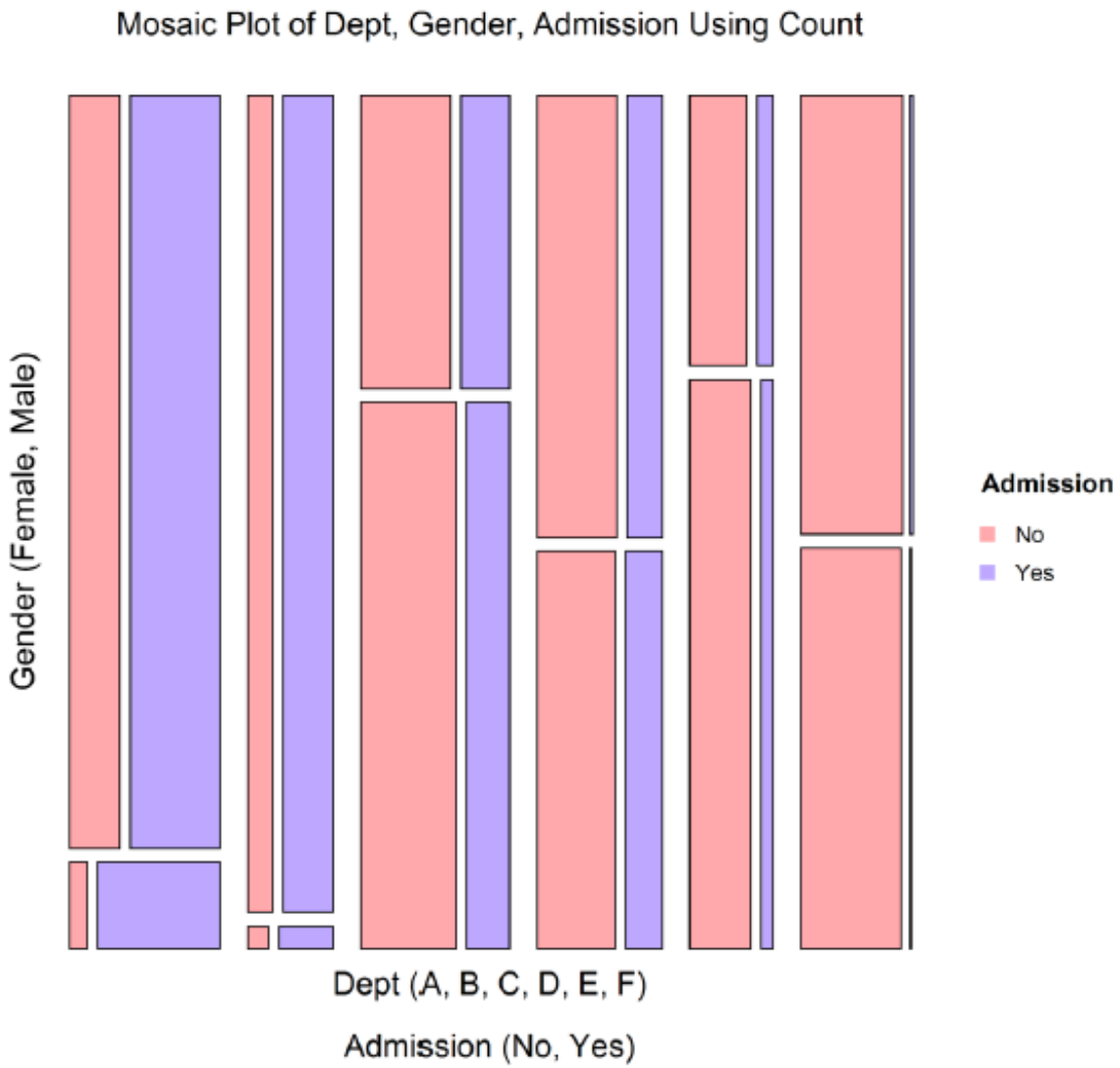
Example : mosaic plot by department



Example : mosaic plot including department

- Departments shown across the bottom
- Percentage of applicants to each department proportional to the width of the bars
- We see that departments A and C have the largest number of applicants and departments B and E have the smallest
- Percent admitted within each gender-by-department combination is width of the corresponding box
- For example, the percentage of females that were admitted to department A (shown by the width of blue box at the lower left) is much larger than that of the males (shown by the width of the long blue box directly above the female box)
- Considering each department in turn by scanning from left to right across the plot, the width of the blue box on the bottom appears to be quite similar to the box directly above it
- This indicates that in most departments the percent of females admitted is *about the same* as that of males admitted

Example : mosaic plot by department

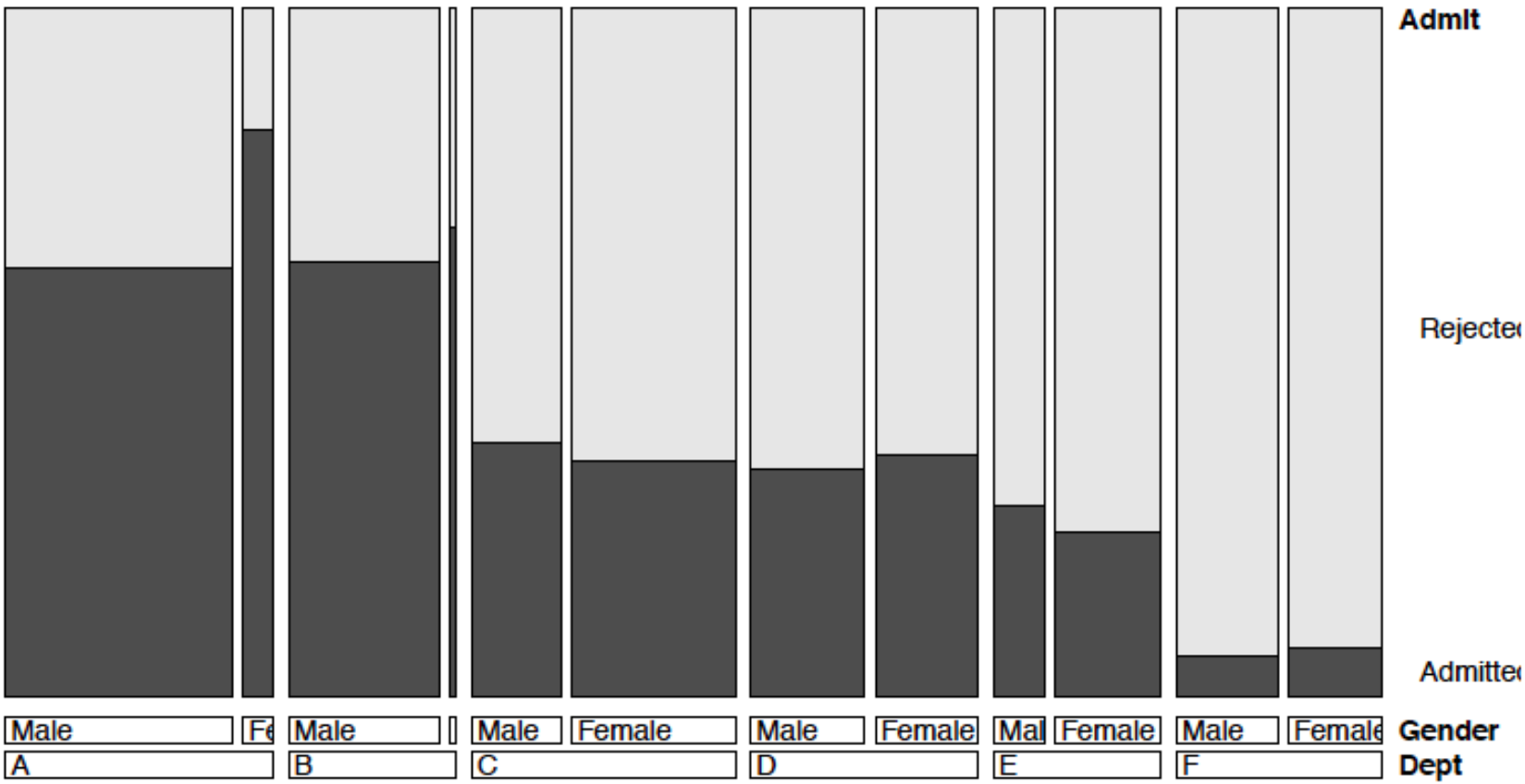


Simpson's paradox

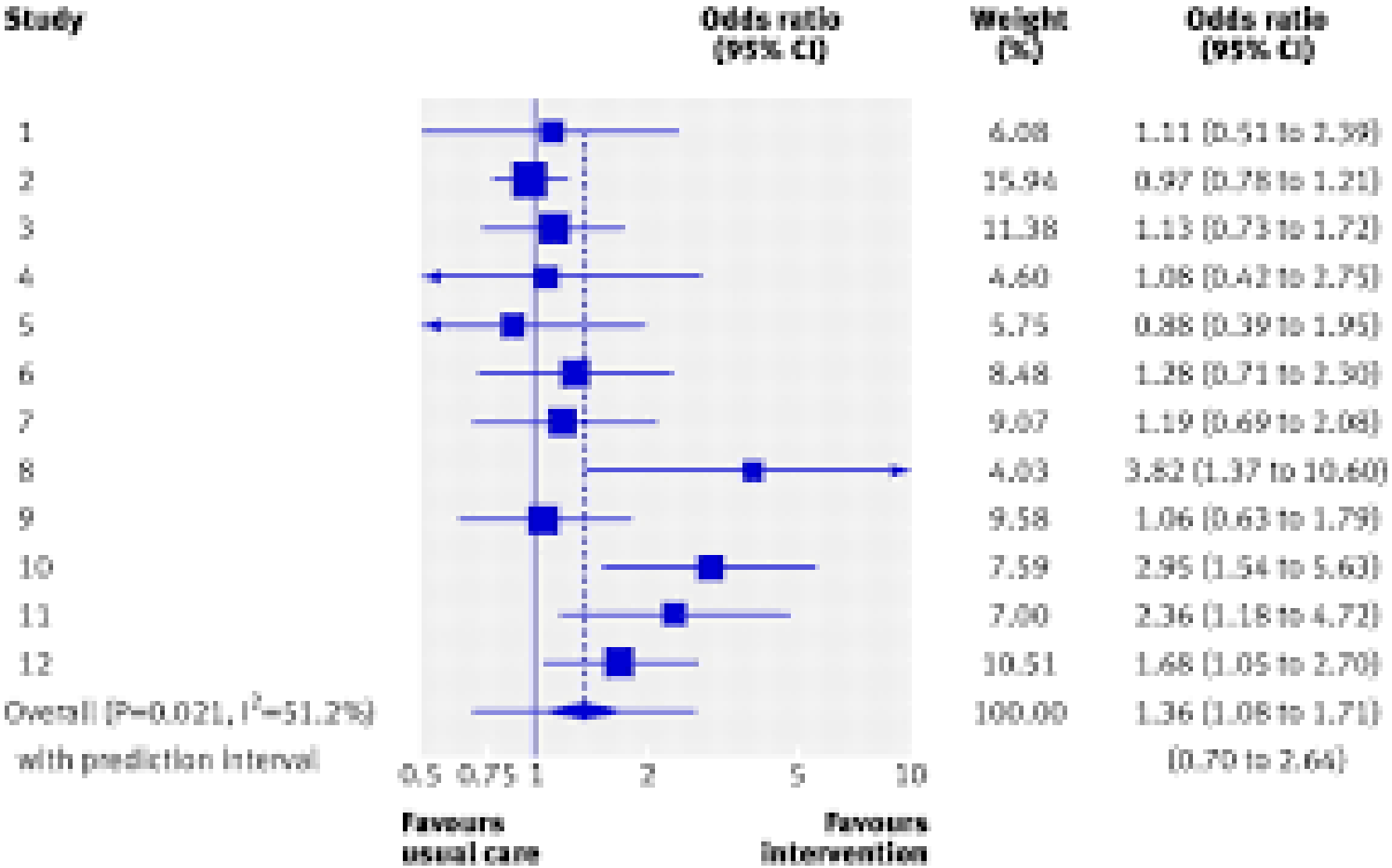
- The tendency of men and women to seek entry to different departments is noticeable
- The research paper by Bickel et al. concluded that women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry)
- This is an example of *Simpson's paradox*, or a *spurious correlation*, where a trend appears in several different groups/strata of data but *disappears or reverses* when these groups are combined

Example : doubledecker plot

```
> doubledecker(Admit ~ Dept + Gender, data=UCBAdmissions[2:1,,])
```



Example : CMH test



Example : Woolf test for 3-way interaction

```
        Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)

data:  UCBAdmissions
X-squared = 17.902, df = 5, p-value = 0.003072
```

Example : fourfold plot