


# Crosstabulation



`img-table-extract.png`

Statistical Reasoning  
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 6

# Outline

---

**Bivariate statistics** look at the probability of an association between two variables and are part of **inferential statistics**.

Before going that far, we need to learn how to **crosstabulate** variables with each other, as tables or graphs.

We will also introduce a few non-parametric tests that do not rely on the normal distribution.

Family Income	Candidate voted for, 1992			
	Clinton %	Bush %	Perot %	Total %
<\$15k	8.3	3.2	2.5	4.7
\$15-30k	10.8	8.4	4.8	8.0
\$30-50k	12.3	11.4	6.3	10.0
\$50-75k	8.0	8.4	3.6	6.7
\$75k+	4.7	6.2	2.1	4.3
<b>Total</b>	8.8	7.5	3.9	6.7

Source: voter.dta

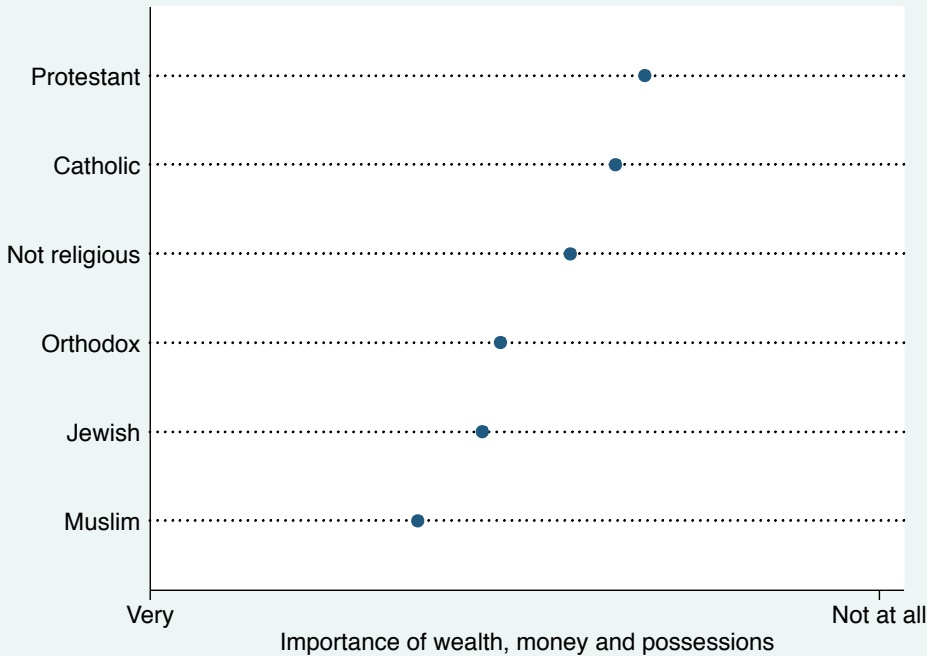
# Visualizations

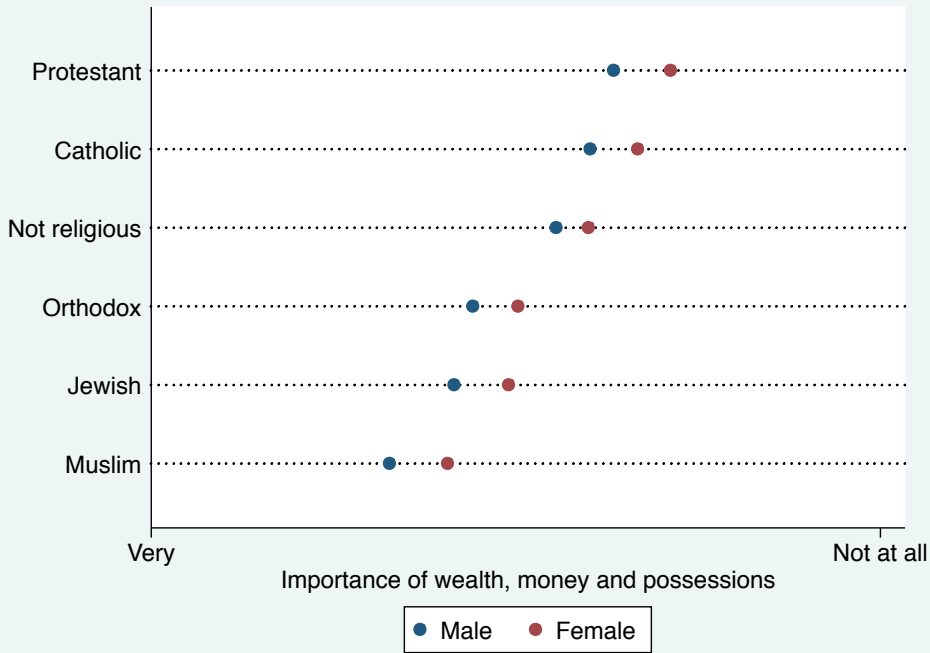
---

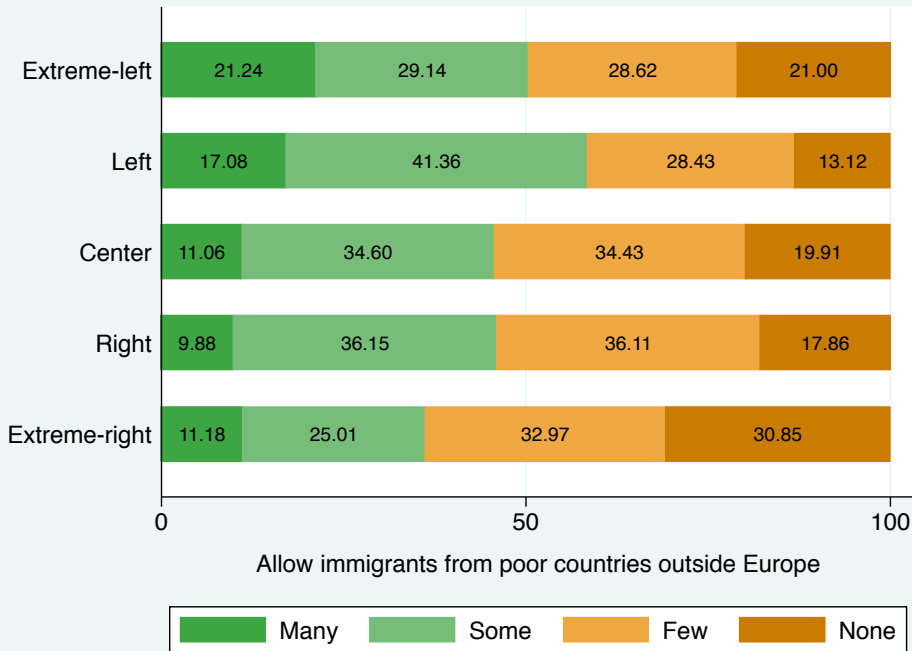
If your independent variables include **categorical variables**, try comparing your dependent variable across their categories:

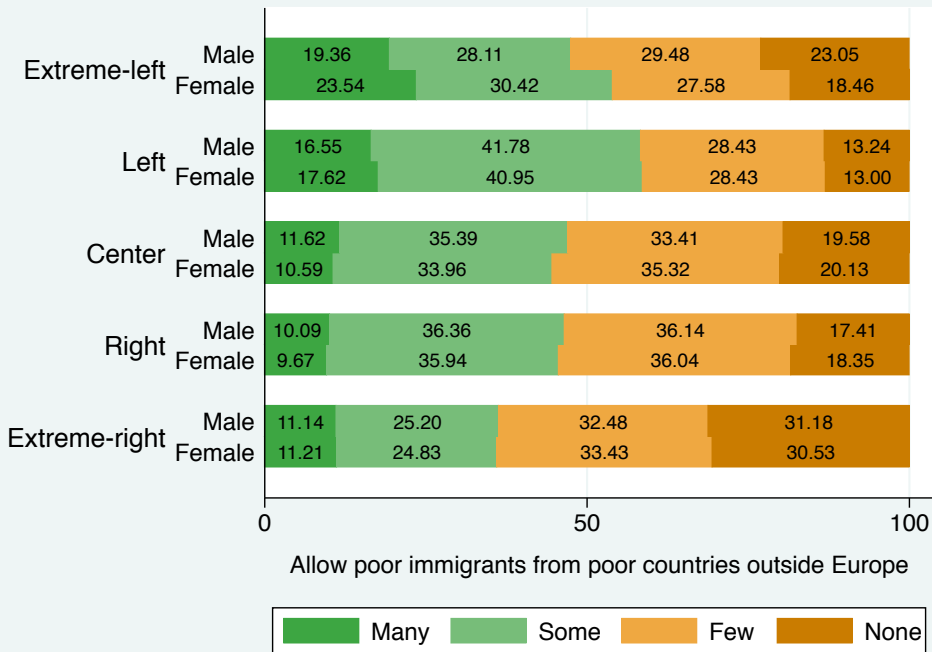
- Some **ordinal scales** can be treated as categories, e.g. high-low income, left-right political position. . .
- Some **nominal variables** are commonly available, e.g. geographical area, religious beliefs, ethnic groups. . .
- Some **binary variables** often act as controls in comparisons, e.g. gender, democratic/dictatorial, religious/atheist. . .

Graphing the variables together in a “two-way graph” explores the **possibility of a relationship** between them.









## Operationalization: Bars and dots

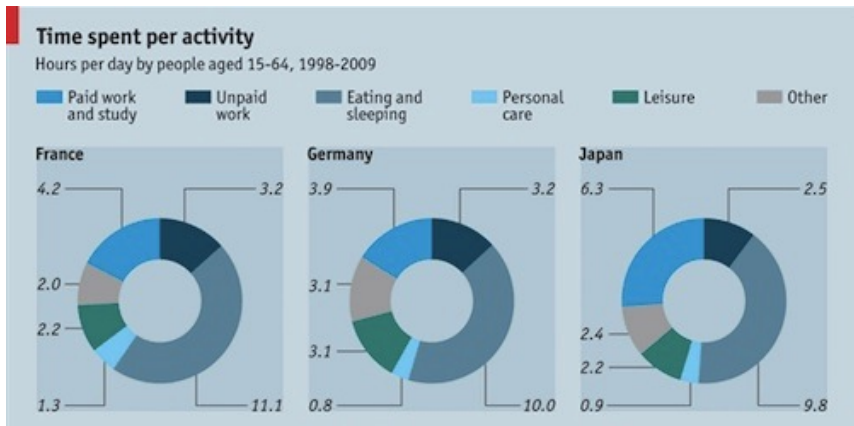
---

- Use `gr dot` to compare the mean, quantile or proportion of a **continuous variable** across groups.
  - Some interval scales can be treated as pseudo-continuous.
  - Use the `sort(1)des` option to order the groups by descending order.
  - Use the `ylab()` option to label the horizontal axis correctly.
  - Use the `exclude0` option if the null value is not meaningful.
- Use `gr bar` or `gr hbar` with the `per stack` options to compare proportions of an **ordinal variable** as percentages across groups.
  - Use `tab, gen()` to create dummies of each group category.
  - Use the `legend(lab(# "..."))` option to rewrite the legending.
  - Use the `bar(,col(...))` option to apply a sensible color scheme.
  - Use the `blab(bar, pos(center) format(%9.2f))` to add labels.
- Play around with `over()`, `by()` and other graph options, looking at course do-files for examples.



# Operationalization: Pies

Technically, you might want to use pie charts with the `by()` option to look at proportions over categories, as they do at *The Economist*:



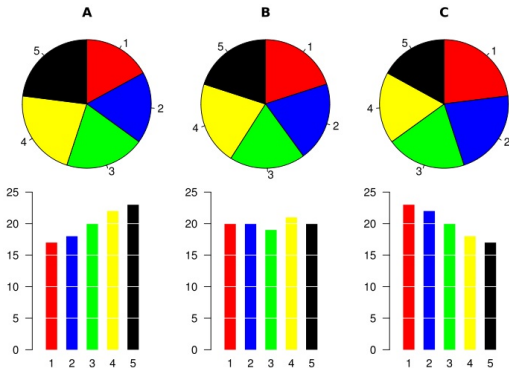
# A Modest Proposal For Pie Charts (Hmm. Pie.)

The problem with pie charts is that the eye reads **polar coordinates** substantially less efficiently than it reads cartesian coordinates.

Simple solution:

**Do not use pie charts.**

(except for rare instances of binary-type proportions shown in exploded slices over a fairly limited number of groups with clear contrasting colours and possibly percentage labels shown outside the plot; even the best pie charts from *The Economist* can be redrawn into more readable plots if you think about it.)



## Crosstabulations

---

Assume a population composed of 50% men and 50% women, with 75% of the population supporting abortion and 25% opposing it.

If men and women hold identical views on abortion, each group will independently reflect the population percentages:

*Example: gender and abortion*

Gender	Views on abortion		
	Support	Oppose	Total
Men	75%	25%	100%
Women	75%	25%	100%

## Crosstabulations

---

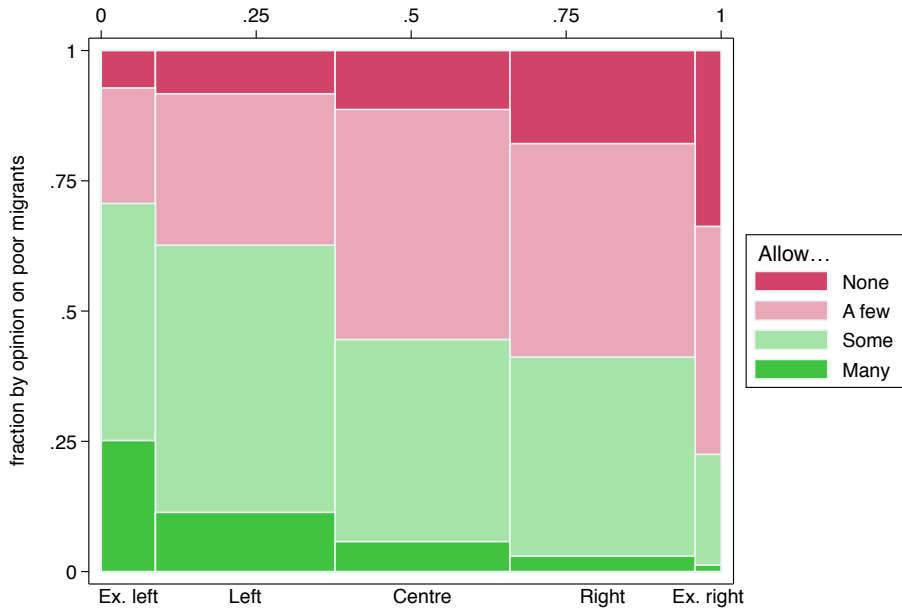
*Example: politics and attitudes towards migrants*

Political categories	Allow poor migrants			
	Many	Some	Few	None
Extreme-left	42	76	37	12
Left	63	284	161	46
Center	31	209	238	61
Right	17	218	234	102
Extreme-right	1	17	35	27

Source: European Social Survey 2008, France.

Note: unweighted data, since the Chi-squared test is non-parametric (it relies on a distinct distribution).

fraction by political scale



## Observed cell percentages

*Example: politics and attitudes towards migrants*

Political categories	Allow poor migrants				Percentage
	Many	Some	Few	None	
Extreme-left	2.20	3.98	1.94	0.63	<b>8.74</b>
Left	3.30	14.86	8.42	2.41	<b>28.99</b>
Center	1.62	10.94	12.45	3.19	<b>28.21</b>
Right	3.46	9.63	9.99	5.54	<b>28.64</b>
Extreme-right	0.98	2.72	2.82	1.56	<b>8.10</b>
<b>Percentage</b>	<b>12.10</b>	<b>33.63</b>	<b>34.90</b>	<b>19.37</b>	<b>100%</b>

The **cell percentage** of each cell in the table is known by comparing each cell frequency to the whole sample distribution.

## Observed row percentages

*Example: politics and attitudes towards migrants*

Political categories	Allow poor migrants				Total
	Many	Some	Few	None	
Extreme-left	25.15	45.51	22.16	7.19	100%
Left	11.37	51.26	29.06	8.30	100%
Center	5.75	38.78	44.16	11.32	100%
Right	2.98	38.18	40.98	17.86	100%
Extreme-right	1.25	21.25	43.75	33.75	100%

The proportion of each row group within each column group is known by reading the **row percentages** of the crosstabulation.

## Observed column percentages

*Example: politics and attitudes towards migrants*

Political categories	Allow poor migrants			
	Many	Some	Few	None
Extreme-left	27.27	9.45	5.25	4.84
Left	40.91	35.32	22.84	18.55
Center	20.13	26.00	33.76	24.60
Right	11.04	27.11	33.19	41.13
Extreme-right	0.65	2.11	4.96	10.89
<b>Total</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

The proportion of each column group within each row group is known by reading the **column percentages** of the crosstabulation.



## Expected frequencies

*Example: politics and attitudes towards migrants*

Political categories	Allow poor migrants			
	Many	Some	Few	None
Extreme-left	13.5	70.3	61.6	21.7
Left	44.6	233.1	204.4	71.9
Center	43.4	226.8	198.8	69.9
Right	46.0	240.2	210.7	74.1
Extreme-right	6.4	33.7	29.5	10.4

The **expected frequency** of each cell is its row total multiplied by its column total, divided by sample size.

## Chi-squared test

---

The **Chi-squared test** works by adding the deviation between observed frequencies  $O_i$  and expected frequencies  $E_i$  for each table cell  $i$ :

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

The values taken by  $\chi^2$  follow the Chi-squared distribution, which provides a probability level for  $H_0 : \chi^2 = 0$  given the degrees of freedom. The test, then, follows a simple logic:

- If  $Pr(\chi^2 = 0) < \alpha$ ,  $H_0$  is rejected: observed frequencies are significantly different from expected ones.
- If  $Pr(\chi^2 = 0) > \alpha$ ,  $H_0$  cannot be rejected: observed frequencies are insignificantly different from expected ones.

## Comparison with odds ratios

---

*"Scotland has 13% redheads, Kabylie has 4% redheads."*

*"Scots are **more likely** to be redheads than Kabyles."*

Quantify the second statement.

- **Treat the dependent variable as a binary 'success/failure':**

1 is success (red hair), 0 is failure (other colour).

$$\text{Odds of } p: \frac{\text{red hair}}{\text{other colour}} = \frac{p}{1-p} = \frac{\text{success}}{\text{failure}}$$

- **Divide the odds in each group to compare across them:** the odds ratio quantifies their comparative likelihood of success.

$$\theta = \frac{\text{odds}_{\text{Scotland}}}{\text{odds}_{\text{Kabylie}}} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\text{success}_1}{\text{failure}_1} \times \frac{\text{failure}_2}{\text{success}_2}$$

## Comparison with odds ratios

*"Scotland has 13% redheads, Kabylie has 4% redheads."*

*"Scots are **more likely** to be redheads than Kabyles."*

Quantify the second statement.

Population	Hair colour	
	Red	Other
Scotland	.13	$1 - .13 = .87$
Kabylie	.04	$1 - .04 = .96$

$$\theta = \frac{.13}{.87} \times \frac{.96}{.04} \approx 3.5$$

Scots are roughly **3.5 times more likely** to have red hair than Kabyles.  
(All figures taken from current Wikipedia estimates.)

## Stata implementation: odds ratio

Are parliamentary regimes **more likely** to select female leaders?

**. logit leader parliamentary, or nolog**

Logistic regression	Number of obs	=	169
	LR chi2(1)	=	0.80
	Prob > chi2	=	0.3698
Log likelihood = -31.809075	Pseudo R2	=	0.0125

leader	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
parliament~y	2.010811	1.51603	0.93	0.354	.4587929	8.81304

Parliamentary regimes are twice more likely to select female leaders: the odds are **substantially large** and yet **statistically insignificant** at  $p < .05$ . Small sample size might have induced a Type II Error.