Course setup

1. Visualizations
2. Crosstabulation
3. Tests

Outline

Equations

$$\text{Linear model: } Y = \beta_0 + beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\text{Logistic model: } Pr(Y = 1) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon)}$$

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.

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...

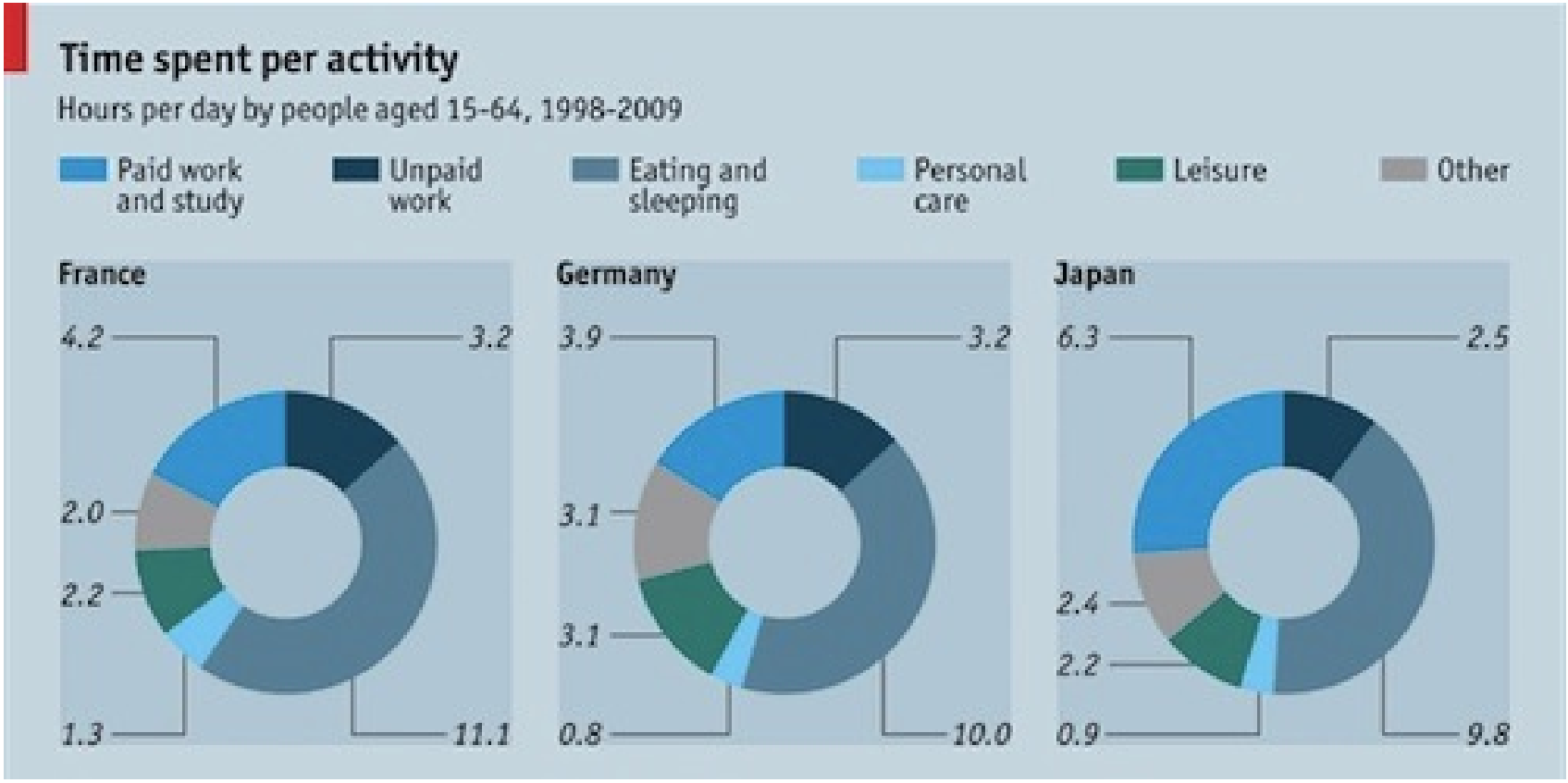
Graphing the variables together in a "two-way graph" explores the possibility of a relationship between them.

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.

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*:

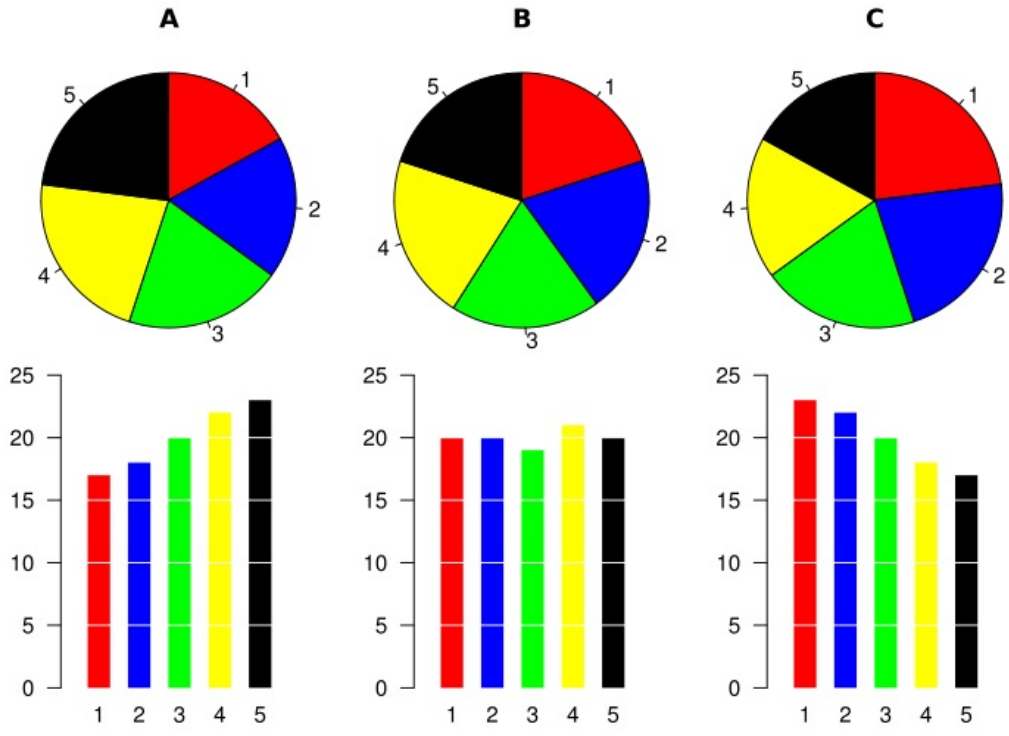


A Modest Proposal For Pie Charts (Hmm. Pie.)

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 plotsif 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:

Crosstabulations

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

Observed cell percentages

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

Observed row percentages

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

Observed column percentages

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

Expected frequencies

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 χ^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).

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

- **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.

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

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.