

Introduction and Fundamentals

MSc Module 6: Introduction to Quantitative Research Methods
Kenneth Benoit

January 7, 2009

Objectives and learning outcomes

- ▶ to introduce a range of methods
- ▶ to identify suitability and purpose of each method
- ▶ conduct and interpret simple analyses
- ▶ to be able to read basic quantitative studies
- ▶ NOT a course in statistics
- ▶ IS primarily introductory and practical

Grading

- ▶ Problem sets (50%)
 - ▶ Handed out Wed, due back the next Wed.
 - ▶ Problem solutions should be submitted as a single pdf file
 - ▶ Problem sets must be submitted to <http://turnitin.com>
 - ▶ You can scan anything using the departmental scanner if needed

- ▶ Final exam (50%). Format is TBA, after last week of class

Texts and Software for this course

- ▶ Two primary texts
 - ▶ Levin, Jack and James Alan Fox. *Elementary Statistics in Social Research: The Essentials*. London: Pearson Education, 2007.
 - ▶ Crawley, Michael J. *Statistics: An Introduction Using R*. Colchester: John Wiley & Sons, Ltd, 2005.
 - ▶ Can be ordered through Amazon or purchased at Hodges and Figgis (purchasing both is definitely recommended)
- ▶ Software will be the R statistical package
 - ▶ Free from <http://www.r-project.org>
 - ▶ Multi-platform
 - ▶ Extremely powerful
 - ▶ Lots of free documentation
 - ▶ There is a GUI shell called R Commander by John Fox, from <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>

A quantitative approach to political analysis

▶ Rationale

- ▶ To allow us to **compare variables** found in the political and social world
- ▶ To **measure** important features of the political and social world
- ▶ To **test hypotheses** about the political and social world
- ▶ To draw **inferences** about the political and social world

▶ Quantity versus quality

- ▶ Quantitative analysis often used to supplement qualitative analyses in political studies
- ▶ Same fundamental approach to knowledge
- ▶ Anything comparable can be quantified
- ▶ Quantity makes analysis reproducible

Prior probabilities and updating

A test is devised to automatically flag racist news stories.

- ▶ 1% of news stories in general have racist messages
- ▶ 80% of racist news stories will be flagged by the test
- ▶ 10% of non-racist stories will also be flagged

We run the test on a new news story, and it is *flagged as racist*.

Question: What is probability that the story is *actually* racist?

Any guesses?

Prior probabilities and updating

- ▶ What about **without the test**?
 - ▶ Imagine we run 1,000 news stories through the test
 - ▶ We expect that 10 will be racist
- ▶ **With the test**, we expect:
 - ▶ Of the 10 found to be racist, 8 should be flagged as racist
 - ▶ Of the 990 non-racist stories, 99 will be wrongly flagged as racist
 - ▶ That's a total of 107 stories flagged as racist
- ▶ So: the **updated** probability of a story being racist, conditional on being flagged as racist, is $\frac{8}{107} = 0.075$
- ▶ The *prior* probability of 0.01 is updated to only 0.075 by the positive test result
- ▶ This is an example of Bayes' Rule,
$$\Pr(R = 1 | T = 1) = \frac{\Pr(T=1|R=1)\Pr(R=1)}{\Pr(T=1)}$$

Some key concepts and terms

Variables These are characteristics of the social and political world that differ from one unit to another.
(Characteristics that do not vary are called *constant*.)

Units of analysis The level or unit at which variables are observed.
Examples: persons, countries, years, country-years.

Hypotheses Statements about the nature of the social and political world, often expressed as statements about relationships between variables.

Measurement Refers to the way in which variables are quantified, and this can take place according to different *levels* of information depending on the nature of the characteristics and how they are observed.

Stages of political research

1. The problem to be studied is reduced to a testable hypothesis.
Example: Incumbents spend more money in campaigns than challengers.
2. An appropriate set of instruments is developed.
Example: A form with categories for reporting is sent to county councils.
3. The data are collected.
Example: Expenditure data is gathered based on candidate disclosures.
4. The data are analyzed for their bearing on the initial hypotheses.
Example: Expenditure data is analyzed conditional on candidate status.
5. Results are interpreted and communicated.
Example: An article is written and published in a journal.

Example

```
> # load in the Dail 2002 candidate spending data
> load("dail2002.Rdata")
> # tabulate incumbency status by victory
> attach(dail2002)
> table(incumbf,wonseatf)
```

| | wonseatf | |
|------------|----------|-----|
| incumbf | Lost | Won |
| Challenger | 266 | 60 |
| Incumbent | 32 | 106 |

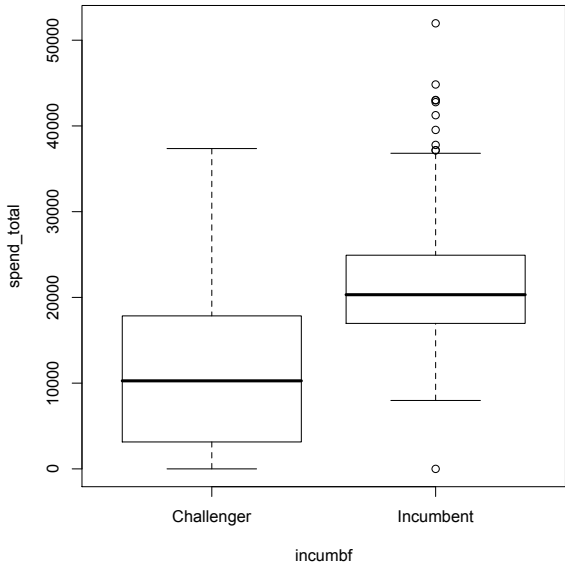
Example (continued)

```
> # produce a mean for each candidate category
> tapply(spend_total, incumbf, mean)
Challenger  Incumbent
  11045.81   21695.70
> # perform a t-test of difference
> t.test(spend_total ~ incumbf)
```

Welch Two Sample t-test

```
data: spend_total by incumbf
t = -12.5634, df = 272.867, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12318.74 -8981.04
sample estimates:
mean in group Challenger  mean in group Incumbent
           11045.81                21695.70

> # plot spending by candidate category
> plot(spend_total ~ incumbf)
```



Levels of measurement

- Nominal** Cases grouped into one and only one category; no relative numerical information even when numerical (e.g. US “ZIP” codes)
- Ordinal** Numbers assigned that rank units – but no indication of the *magnitude* of differences
- Interval** Numbers indicate *exact difference* between units and use *constant units of measurement*
- Ratio** Ratios between measurements as well as intervals are meaningful because there is a starting point (zero)

Examples of variables with different levels of measurement

- ▶ Age
- ▶ Religious denomination
- ▶ Temperature
- ▶ Party preference
- ▶ Level of education
- ▶ GDP per capita
- ▶ Gender
- ▶ Strongly disagree, disagree, no opinion, agree, strongly agree
- ▶ Campaign spending in euros

Introduction to Data

- ▶ The difference between tables and *datasets*
- ▶ This is a **table**:

| | Lost | Won |
|------------|------|-----|
| Challenger | 266 | 60 |
| Incumbent | 32 | 106 |

- ▶ This is a (partial) **dataset**:

| | district | incumbf | wonseatf |
|-----|--------------------|------------|----------|
| 1 | Carlow Kilkenny | Challenger | Lost |
| 2 | Carlow Kilkenny | Challenger | Lost |
| 5 | Carlow Kilkenny | Incumbent | Won |
| 100 | Donegal South West | Challenger | Lost |
| 459 | Wicklow | Incumbent | Won |
| 464 | Wicklow | Challenger | Lost |

- ▶ How computers record and represent data is important
- ▶ A variety of tools exist (computer-wise)