

Introduction and Dealing with Data

PO7001: Quantitative Methods I
Kenneth Benoit

September 29, 2010

Objectives and learning outcomes

- ▶ Understand data concepts and basic descriptive quantitative analysis tools
- ▶ Work with real datasets to perform basic quantitative analyses
- ▶ Graph data effectively for presentation and analysis
- ▶ Recognize and understand the basics of the linear regression model
- ▶ Use the R statistical software package for analyzing and graphing data
- ▶ Understand sufficient theoretical and practical material to build on in a second, more advanced quantitative methods course

Grading

- ▶ Problem sets (50%)
 - ▶ Handed out Wed, due back the next Wed, every other week
 - ▶ So a total of 5 problem sets, each 10%
 - ▶ 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
- ▶ Course paper (50%) – see guidelines on course handout
- ▶ No (final) exam

Texts for this course

- ▶ One primary text:
Verzani, John. 2005. Using R for introductory statistics. Boca Raton, FL: Chapman & Hall/CRC.
- ▶ One recommended, companion text:
Crawley, Michael J. Statistics: An Introduction Using R. Colchester: John Wiley & Sons, Ltd, 2005.
- ▶ One highly recommended additional text:
Gelman, Andrew and Jennifer Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Software for this course: R

- ▶ *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

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.

A quick R example

```
# show what we can do with R
packages    <- c("SPSS","Excel","Stata","R")
features    <- c(-50,50,100,150)
cost        <- c(100,50,90,0)
barplot(features, names.arg=packages, ylab="Feature index")
plot(features, -cost, xlab="Feature Index",
      ylab="(Inverse) Cost Index")
text(features, -cost, packages, pos=c(4,4,4,2))
```

The Birthday Problem

- ▶ The “birthday problem” is a classic problem: Given a group of people, what is the probability of two people in the group having the same birthday?
- ▶ Approach problems like this first by considering extreme cases:
 - ▶ With one person, $\Pr()=0$
 - ▶ With two persons, probability is very small
 - ▶ For a group of ≥ 365 people, $\Pr()=1$ that two people will have the same birthday, since there are only 365 possible birthdays to go around
- ▶ Question: How many people are needed before the probability exceeds 0.50?

- ▶ For one person, there are 365 distinct birthdays (excluding leap years for simplicity)
- ▶ For two people:
 - ▶ there are 365^2 different pairs of birthdays for our two people
 - ▶ there are $(365-1)$ different ways for the second person to have a different birthday
 - ▶ so the $\Pr(x_1 \neq x_2) = \frac{365(365-1)}{365^2}$
- ▶ A third person has 363 different days that could be different from the first two people, so
 $\Pr(x_1 \neq x_2 \neq x_3) = (365 \cdot 364 \cdot 363)/365^3$
- ▶ This leads to the general formula for the probability of a match with n birthdays being

$$\begin{aligned}
 Pr(n) &= 1 - \frac{365^n - 365 \cdot 364 \cdot \dots \cdot (365 - n + 1)}{365^n} \\
 &= 1 - \frac{365!}{(365 - n)!365^n}
 \end{aligned}
 \tag{1}$$

A more extended R 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

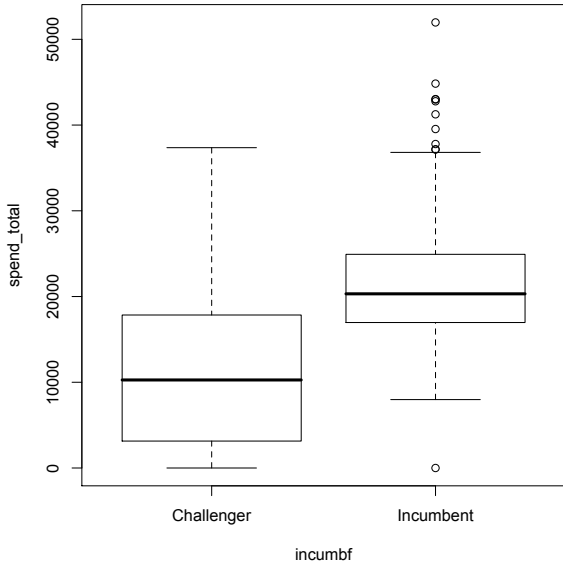
Example (continued)

```
> # 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)
```



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)

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)

Working with R

- ▶ Unlike (e.g.) SPSS, R uses a command line as its main interface
- ▶ R can perform base arithmetic (“R as a calculator”)
- ▶ R has a large number of built-in functions
- ▶ Functions have required arguments, optional arguments, named arguments
- ▶ R can be *programmed* to add additional functions
- ▶ *Assignment* is done with `<-`
- ▶ `c()` is used to enter data, separated by `,`

R and objects

- ▶ Everything in R is an *object*
- ▶ Every object has a **mode**
 - ▶ “Atomic modes”: logical, numeric, complex, character, raw, NA
 - ▶ lists
 - ▶ functions
 - ▶ expressions
- ▶ Every object has a **length**
- ▶ Every object also has a **class**, e.g. "numeric", "logical", "character", "list", "matrix", "array", "factor", "data.frame"

Data types in R

Vectors are atomic structures containing components of all the same mode (e.g. a numeric vector, or a character vector)

Matrixes/Arrays are multi-dimensional generalizations of vectors that can be indexed by two or more indices, and are printed in special ways

Factors provide compact ways to handle categorical data.

Lists are a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists

Data frames are matrix-like structures, in which the columns can be of different types; usually these are “datasets” with one row per observation, and columns as variables

Functions code objects that take arguments and return values

Working in R: Birthday problem example

- ▶ Formula: $1 - \frac{365!}{(365-n)!365^n}$
- ▶ In R, we can use the `factorial()` function
- ▶ So for $n = 10$:
`1 - (factorial(365) / (factorial(365-n) * 365^n))`
- ▶ Does this work? **No – numbers too big!** ($63! \approx 1.982 \times 10^{87}$ – the upper range of current estimates of the number of known particles in the universe)
- ▶ **So what can we do???**
- ▶ How to solve this: use *logarithms* – and the log factorial function, `lfactorial()`
`1-exp(lfactorial(365) - lfactorial(365-n) - n*log(365))`

Working in R: Birthday problem example code

```
lbdp <- function(n) {  
  1 - exp(lfactorial(365) - lfactorial(365-n) - n*log(365))  
}  
  
x <- 1:60  
plot(x,lbdp(x))  
  
plot(x,lbdp(x),  
      xlab="Number of people",ylab="Probability of same birthday")  
abline(h=.5, lty="dashed", col="red")
```

