Discrete Choice Modelling

Introduction

Iragaël Joly

Grenoble-Inp, Génie Industriel - Gael

iragael.joly@grenoble-inp.fr

October 2017

Table of contents

Introduction

Program

A discret choice course

- From the econometrics of qualitative variables to the econometrics of discrete choices
- From theoretical models of behaviours to applied model with `R` software
- 5 steps/lessons :
  - Introduction
  - Binary Choice
  - Multinomial Choices
  - Advanced Models : papers presentation (with application)
  - Discussion : papers discussion (reversed class)
  - Guided work sessions

Motivation

Human dimension

- Demand concept in engineering, business, marketing, planning, policy making
- Operational in many fields (transportation, marketing, energy, finance, etc.)
- Choice of product, brand, mode, destination, contract type and usage, buy and sell, etc.
- Need for behavioral theories, quantitative methods, operational mathematical models

Typical choice questions

McFadden (2014) merges the three elements of consumption choices :

Three choice questions

"Which" : Product choice : option choice
mode choice, itinary choice, etc.

"How many" : Product quantity
Number of trips, motorisation rate, travelled distance
travel time

"When" : Moment or length of a consumption
Car replacement, departure time, duration between events

Econometrics

Statistical and Econometric Models

- How much & When : Duration model (survival model)
- How many : Count models
- Which : Discrete Choice Models

This course

Focus

- Microeconometrics perspective
- To analyse individual behavior (vs. aggregate behavior)
- To present theory of behavior which is
  - descriptive (how people behave) and not normative (how they should behave)
  - general : not too specific
  - operational : can be used in practice for forecasting, value elicitation, etc
- Type of behavior : modeling discrete outcomes
  "Which one", not "When ?", not "How many ?"
- Dependent variable is not a quantitative measure
- Modeling probabilities of events occurence
  rather than conditional mean functions

Importance of DCM in Economics

Daniel L. McFadden

- UC Berkeley 1963, MIT 1977, UC Berkeley 1991
- Laureate of The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2000

Course Objectives

- Theory : Theoretical underpinnings of the models
  - Models and data generating mechanisms
  - Econometric Theory
- Practice
  - Tools for estimation and inference
  - Model specification
  - Analysis and interpretation of numerical results

General Context of Choice analysis

General context of Choice analysis

General context

- Random Utility view of the choices that are observed.
- The decision maker is faced with a set of alternatives
- She reveals underlying preferences by the choice that are made
- Choices made will be affected by observable influences (ex : advertising) and unobservable characteristics of the chooser (ex : mood).

Four broad frameworks

Binary Choice : A choice between a pair of options

- taking or not an action
- decision be between two distinctly different choices, (public or private transportation).
- the 0/1 outcome is a label for "no/yes"

Multinomial Choice : A choice among more than two choices

- the observed response is simply a label (a brand, a place, the travel mode)
- numerical assignments are not meaningful

Four broad frameworks

Ordered Choice : The individual reveals the strength of his or her preferences with respect to a single outcome.

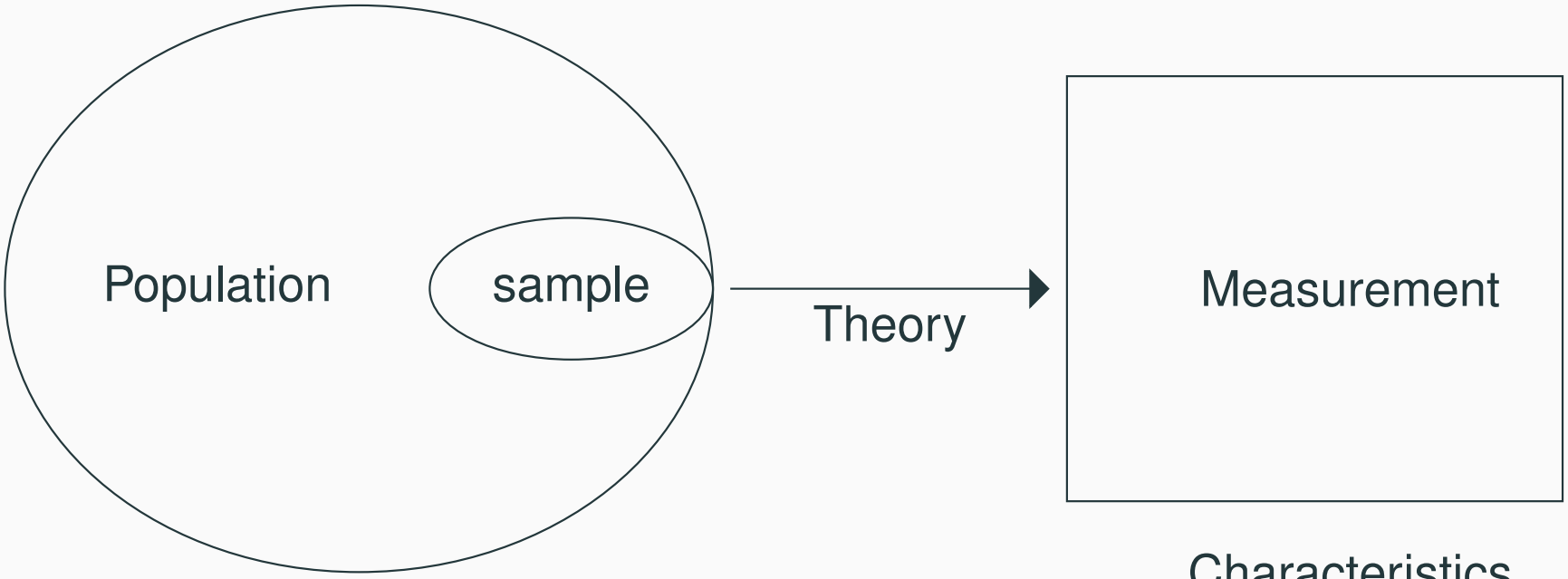
- Familiar cases : strength of feelings about a commodity or self-assessed well-being.
- Opinions are given meaningful numeric values, usually 0, 1, . . . , J for some upper limit, J .
- numerical values are only a ranking, not a quantitative measure
- Thus a "1" is greater than a "0" in a qualitative sense, but not by one unit,
- The difference between a "2" and a "1" is not the same as that between a "1" and a "0."

Four broad frameworks

Event Counts : The observed outcome is a count of the number of occurrences.

- Similar to the preceding three settings : "dependent variable" measures an individual choice (number of visits...)
- The event count might be the outcome of some natural process, such as incidence of a disease in a population (or the number of defects per unit of time in a production process).
- The models will still be constructed to accommodate the discrete nature of the observed response variable.

Sample and Measurement



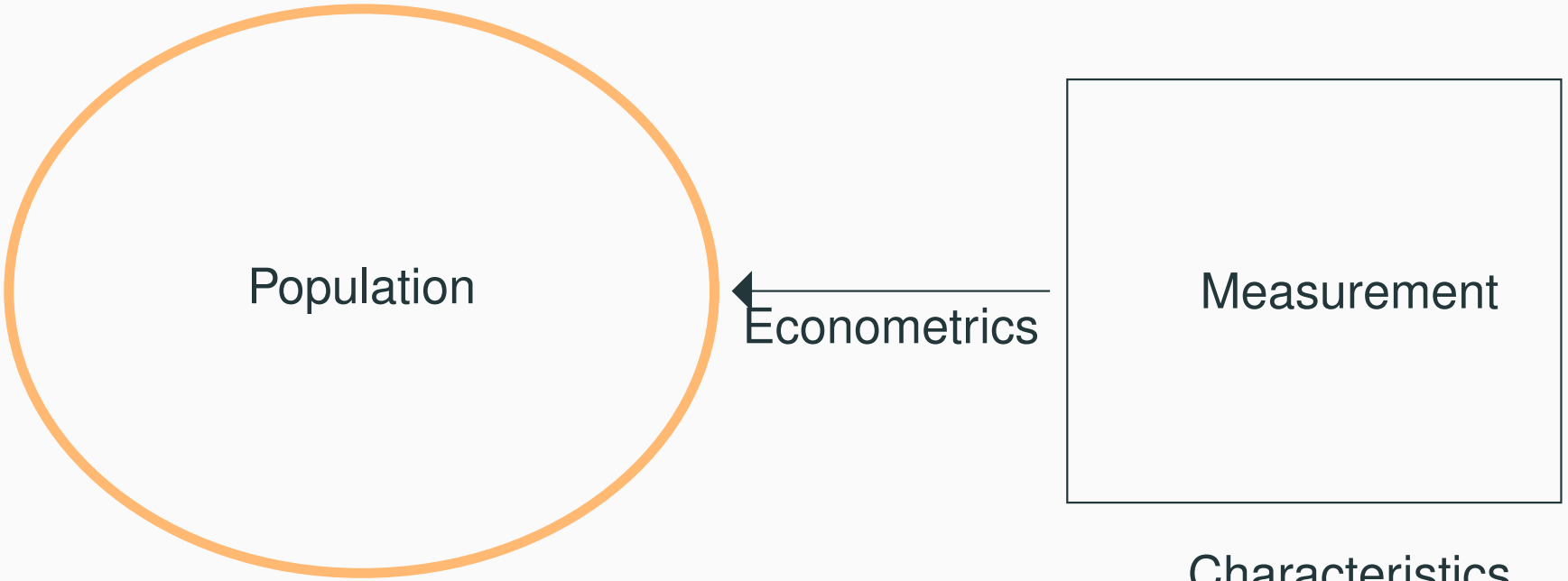
Characteristics

Behavior

Patterns

Choices

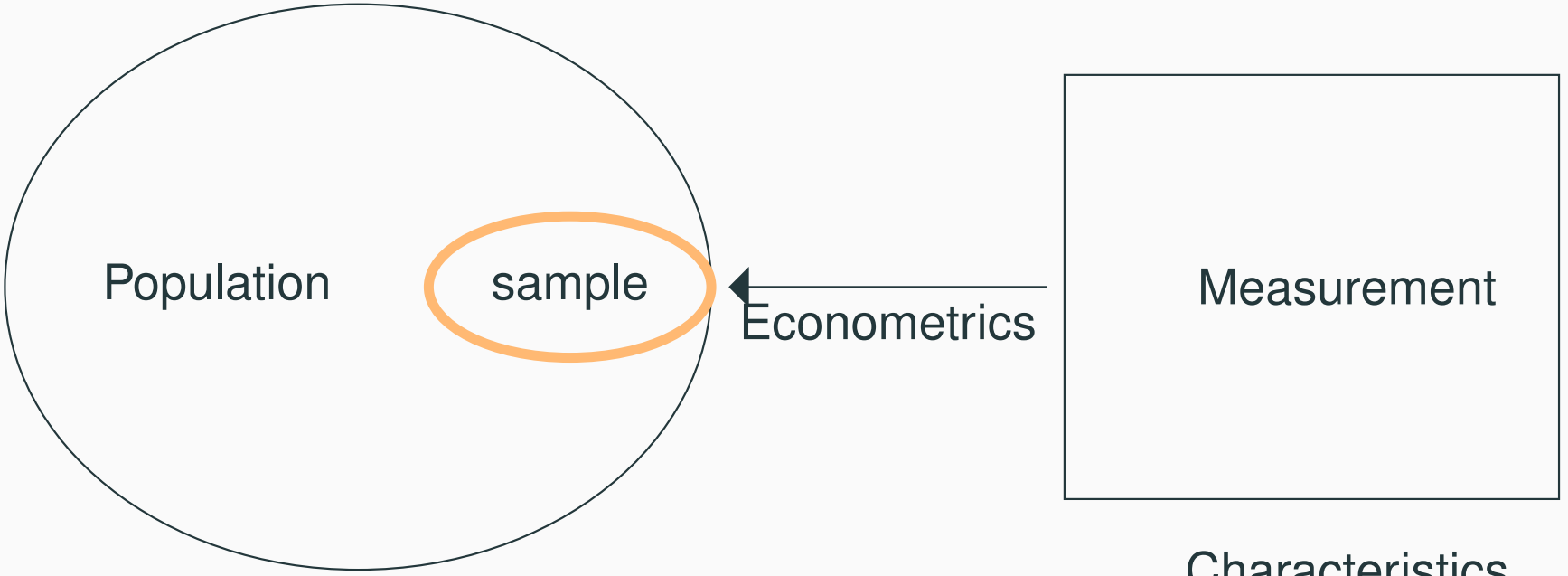
Classical Inference



Imprecise inference
sampling theory
and asymptotics

Characteristics
Behavior
Patterns
Choices

Bayesian Inference



'exact' inference
about only the sample
the 'posterior' density

Characteristics
Behavior
Patterns
Choices

Issues in Model Building and Regression Basics

- Model : the conditional mean function
  - modeling the mean
  - modeling the probabilities for DCM
- Estimation
  - Coefficients
  - Interesting Partial Effects
- Functional Form and Specification
- Statistical Inference
- Prediction
  - Individuals
  - Aggregates
- Model Assessment and Evaluation

R and Tests for Categorical Data

Introduction

Proportion Tests - Motivation

- Observation of choices and categorical data
- Theory of test and application with `R`
- Statistical inference and indicators of interest : confidence interval, p-value, and odd ratio
- What are the associated research questions ?

Inference for a (Single) Proportion (1)

Proportion Test

- $H_0 : \pi = 0.5$ is tested against the two-sided alternative $H_1 : \pi \neq 0.5$
- a 95% confidence interval for π is calculated
- both the test and the CI incorporate a continuity correction.

Example

Half of the students have experienced inferencial test with `R`?
Our sample : $n = 20$, `R` users :$n_1 = 8$

```
prop.test(8,20,p=.5,alternative="two.sided",
    conf.level=0.95,correct=TRUE)
```

Inference for a (Single) Proportion (2)

```
##
##  1-sample proportions test with continuity correction
##
## data:  8 out of 20, null probability 0.5
## X-squared = 0.45, df = 1, p-value = 0.5023
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1997709 0.6358833
## sample estimates:
##   p
## 0.4
```

Inference for a (Single) Proportion (1)

Proportion Test - Exercise

Explain this new test and conclude ?

```
prop.test(8,20,p=.4,alternative="greater",
          conf.level=0.99,correct=FALSE)
```

Inference for a (Single) Proportion (2)

```
##
##  1-sample proportions test without continuity correction
##
## data:  8 out of 20, null probability 0.4
## X-squared = 0, df = 1, p-value = 0.5
## alternative hypothesis: true p is greater than 0.4
## 99 percent confidence interval:
##  0.194216 1.000000
## sample estimates:
##   p
## 0.4
```

Two-Way Contingency Tables (1)

The table as Matrix

```
Ruser <- matrix(c(435,147,375,134),nrow=2,byrow=TRUE)
dimnames(Ruser) <- list(c("Female","Male"),c("Yes","No"))
names(dimnames(Ruser)) <- c("Gender","R.User")
#addmargins(Ruser)
```

```
##          R.User
## Gender    Yes  No
##   Female 435 147
##   Male   375 134
```

Two-Way Contingency Tables (2)

```
round(Ruser / sum(Ruser), 3)
rowtot <- apply(Ruser,1,sum)
coltot <- apply(Ruser,2,sum)
sweep(Ruser,1,rowtot,"/") #apply "/" to Ruser by row
sweep(Ruser,2,coltot,"/")
```

Two-Way Contingency Tables (3)

```
##          R.User
## Gender      Yes    No
##   Female 0.399 0.135
##   Male   0.344 0.123
##          R.User
## Gender          Yes        No
##   Female 0.7474227 0.2525773
##   Male   0.7367387 0.2632613
##          R.User
## Gender         Yes        No
##   Female 0.537037 0.5231317
##   Male   0.462963 0.4768683
```

Two-Way Contingency Tables (1)

The table as Dataframe

```
Gender <- c("Female","Female","Male","Male")
R.user <- c("Yes","No","Yes","No")
Count <- c(435,147,375,134)
Ruserdf <- data.frame(Gender,R.user,Count)
Ruserdf
rm(Gender, R.user, Count)
names(dimnames(Ruserdf)) <- c("Gender","R user")
tab1 <- tapply(Ruserdf$Count,list(Ruserdf$Gender,
                          Ruserdf$R.user),c);tab1
```

Two-Way Contingency Tables (2)

```
##   Gender R.user Count
## 1 Female    Yes   435
## 2 Female     No   147
## 3   Male    Yes   375
## 4   Male     No   134
##         No Yes
## Female 147 435
## Male   134 375
```

Two-Way Contingency Tables (1)

The test on matrix

```
prop.test(Ruser)
```

Two-Way Contingency Tables (2)

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  Ruser
## X-squared = 0.11103, df = 1, p-value = 0.739
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04321353  0.06458148
## sample estimates:
##    prop 1    prop 2
## 0.7474227 0.7367387
```

Two-Way Contingency Tables (1)

The test on Dataframe

```
prop.test(tab1)
```

Two-Way Contingency Tables (2)

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tab1
## X-squared = 0.11103, df = 1, p-value = 0.739
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06458148  0.04321353
## sample estimates:
##    prop 1    prop 2
## 0.2525773 0.2632613
```

Comparing Proportions (1)

Aim

Identify *significant* factor effect on a categorical variable (or discrete choice)

Let take an example and define parameters for comparing groups

difference, relative risk, odds ratio

Comparing Proportions (2)

	Disease Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

p_1 : probability of success (disease) in row 1

$1 - p_1$: probability of failure (no disease) in row 1

p_2 : probability of success (disease) in row 2

$1 - p_2$: probability of failure (no disease) in row 2

Comparing Proportions (3)

Difference of proportions

The difference of proportions of successes : $p_1 - p_2$ (basic comparison of the two rows)

Comparison on failures is equivalent : $(1 - p_1) - (1 - p_2) = p_1 - p_2$

- between -1 and +1
- equals zero when identical conditional distribution : Y is independant
- may have greater importance when both p_i are close to 0 or 1

Comparing Proportions (4)

Relative Risk

The relative risk is the ratio of probabilities

relative risk : $= \frac{p_1}{p_2}$

- nonnegative real number
- A relative risk of 1.0 corresponds to independence

Comparing Proportions (5)

Let define the odds

The odds of getting disease for the people who were exposed to the risk factor (Risk factor = Yes (1)) :

	Disease	Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

p_1 : probability of success (disease) in row 1

$1 - p_1$: probability of failure (no disease) in row 1

$$O_X = \frac{p_1}{1 - p_1} \sim \frac{\hat{p_1}}{1 - \hat{p_1}} = \frac{50/70}{20/70} = 2.5$$

Comparing Proportions (6)

The odds of getting disease for the people who were not exposed to the risk factor (Risk factor = No (0))

	Disease	Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

p_2 : probability of success (disease) in row 2

$1 - p_2$: probability of failure (no disease) in row 2

$$O_{NoX} = \frac{p_2}{1 - p_2} \sim \frac{\hat{p}_2}{1 - \hat{p}_2} = \frac{100/230}{130/230} = 0.77$$

The Odds-Ratio (1)

The Odds Ratio of having disease for people who were exposed to the risk factor versus not exposed

$$\theta = OR = \frac{O_X}{O_{NoX}} \sim \frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = \frac{50 \times 130}{20 \times 100} = 3.25$$

Interpretation : The odds of having disease are 3.25 times higher for those who were exposed to the risk factor than those who were not exposed to the risk factor.

The Odds-Ratio (2)

- If $\theta > 1$, then the odds of success are higher for exposed individuals than for not exposed
- If $\theta < 1$, then the odds of success are lower for exposed individuals than for not exposed
- If $\theta = 1$, then the odds of success are equal for exposed individuals than for not exposed

- For inference : independence corresponds to $log(\theta) = 0$
- The log odds ratio is symmetric about this value

OR, RR, DR ? (1)

Limits of DR :

- DR cannot be constant
- DR depend on the initial risk

Limits of RR :

- RR have limited range of risk
- ex. $RR = 2$ can only apply to risks below .5
- above that point the RR must diminish
- Risk ratios are similar to odds ratios if the risk is small

OR, RR, DR ? (1)

OR is a natural description of an effect in a probability model since an OR can be constant.

Without Risk Factor		With Risk Factor	
Probability	Odds	Odds	Probability
Risk Factor	Yes (1)	No (0)	Total
.2	.25	.5	.33
.5	1	2	.67
.8	4	8	.89
.9	9	18	.95
.98	49	98	.99

Effect of an odds ratio of 2 on various risks

OR, RR, DR ? (1)

- OR : unlimited range $\rightarrow$ can describe an effect over the entire range of risk
- positive OR still yield valid probability
- constant OR describe the effect independently of other covariables

OR, RR, DR ? (2)

Let X_1 a binary factor and $A = \{X_2, \ldots, X_p\}$ other factors (independant from X_1)

The estimate of $DR = P(Y = 1 | X_1 = 1, A) - P(Y = 1 | X_1 = 0, A)$ is :

$$
\begin{aligned}
DR = & \frac{1}{1 + exp\{-(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots)\}} \\
& - \frac{1}{1 + exp\{-(\hat{\beta}_0 + \hat{\beta}_2 X_2 + \ldots)\}} \\
= & \frac{1}{1 + \left(\frac{1-\hat{R}}{\hat{R}}\right) exp(-\hat{\beta}_1)} - \hat{R}
\end{aligned}
$$

OR, RR, DR ? (3)

The DR estimate can be plotted against $\hat{R}$ or against levels of A to display absolute risk increase against overall risk.

Fig. 1 depicts the relationship between risk of a subject without the risk factor and the increase in risk for a variety of relative increases (odds ratios). It demonstrates how absolute risk increase is a function of the baseline risk.

OR, RR, DR ? (4)

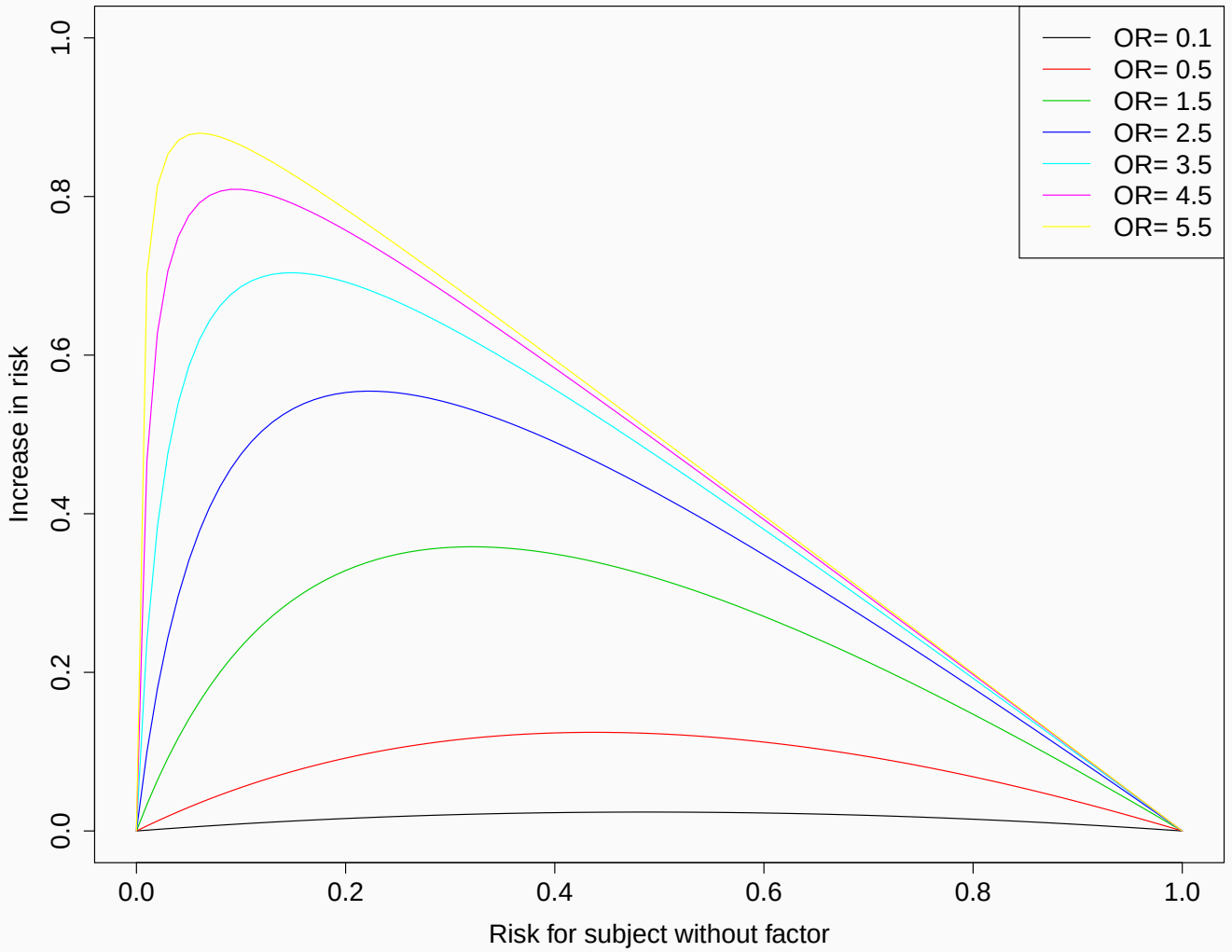


FIGURE 1 – Absolute benefit as a function of the baseline risk. The odds ratios are given for each curve

The Odds-Ratio with R (1)

Go back on our example of R users

The Odds-Ratio with R (2)

```
tab1
```

```
##          No Yes
## Female 147 435
## Male   134 375
```

```
R.test <- prop.test(tab1)
names(R.test)
```

```
## [1] "statistic"   "parameter"   "p.value"     "estimate"    "null.valu
## [6] "conf.int"    "alternative" "method"      "data.name"
```

```
p1 <- 147/(147+435); p2 <- 134/(134+375)
p1;p2
```

```
## [1] 0.2525773
## [1] 0.2632613
```

The Odds-Ratio with R (1)

Proportion of "No"

```
R.test$estimate
```

```
##    prop 1    prop 2
## 0.2525773 0.2632613
```

The Odds-Ratio with R (1)

Odds of "No" over "Yes"

```
odds <- R.test$estimate/(1-R.test$estimate)
names(odds) <- c("Odds1: F","Odds2: M")  ; odds

##  Odds1: F  Odds2: M
## 0.3379310 0.3573333
```

Odds are almost equivalent

The Odds-Ratio with R (1)

Odds ratio

```
OR <- odds[1]/odds[2]
names(OR) <- c("OR")  ; OR

##        OR
## 0.9457025
```

The odds of success (not R-user) are "lower" for exposed individuals (women) than for not exposed (men)

Confidence Interval for Odds Ratio (1)

For large sample, the log of odds ratio, $ln(\hat{\theta})$, follows asymptotically a normal distribution.
The $(1-\alpha)100\%$ confidence interval estimate for the Log Odds Ratio is

$$ln(\hat{\theta}) \pm z_{\alpha/2}\hat{s}$$

The $(1-\alpha)100\%$ confidence interval estimate for the Odds Ratio is

$$\left(e^{ln(\hat{\theta})-z_{\alpha/2}\hat{s}}, e^{ln(\hat{\theta})+z_{\alpha/2}\hat{s}}\right)$$

where $\hat{s} = \sqrt{\sum(1/n_{ij})}$ (from Delta Method)

Confidence Interval for Odds Ratio (1)

CI for Odds Ratio

```
theta <- odds[1]/odds[2]
ASE <- sqrt(sum(1/tab1))
# ASE
ASE
logtheta.CI <- log(theta) + c(-1,1)*1.96*ASE
# IC log(theta)
logtheta.CI
# IC(OR)
exp(logtheta.CI)
```

```
## [1] 0.1386756
## [1] -0.3276314  0.2159770
## [1] 0.7206286 1.2410738
```

Confidence Interval for Odds Ratio (2)

Interpretation ?

The Odds-Ratio (1)

Automatic calculus with function `oddsratio.R`

```
source(file = "Rscripts\\oddsratio.R")
odds.ratio(tab1)
```

The Odds-Ratio (2)

```
## $estimator
## [1] 0.9457025
##
## $ASE
## [1] 0.1386756
##
## $conf.interval
## [1] 0.7206322 1.2410676
##
## $conf.level
## [1] 0.95
```

The Odds-Ratio (1)

Exercise : Seat-Belt Use and Traffic Deaths ***(Agresti, 2013)***

Fatality results for children under age 18 who were passengers in auto accidents in Florida in 2008 according to whether the child was wearing a seat belt

	Injury	Outcome	
Seat-Belt Use	Fatal	Nonfatal	Total
No	54	10,325	10,379
Yes	25	51,790	51,815

The Odds-Ratio (2)

Solution

- The sample odds ratio $\hat{\theta} = 10.83$
- Estimated standard error $\hat{s}$ of $log(\hat{\theta}) = 2.383$ is $\hat{s}(log(\hat{\theta})) = 0.242$
- A 95% confidence interval for $log(\hat{\theta})$ is $2.383 \pm 1.96(0.242)$, or $(1.908, 2.857)$.
- interval fore is $[exp(1.908), exp(2.857)]$ or $(6.74, 17.42)$
- There is a very strong association
- Even though the overall sample size is extremely large, the estimate of the true odds ratio is rather imprecise because of the relatively small number of fatalities

Chi-Squared Tests of Independence (1)

In the `MASS` package : dataframe `survey` :

`Smoke` records the students smoking habit : "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never"

`Exer` records their exercise level : "Freq" (frequently), "Some" and "None".

Question : Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```
library(MASS) ;  library(gplots)
tbl = table(survey$Smoke, survey$Exer)
tbl
balloonplot(t(tbl), main ="Smokers and Exercise"
        ,xlab ="Exercise", ylab="Smoke"
        , label = FALSE, show.margins = FALSE)
```

Chi-Squared Tests of Independence (2)

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

Chi-Squared Tests of Independence (3)

Chi-Squared Tests of Independence (1)

The Pearson statistic :

$$\Delta^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

Δ^2 is asymptotically $\chi^2(I-1) \times (J-1)$

Chi-Squared Tests of Independence (1)

Chi-Squared test

```
chisq.test(tbl)
```

```
## Warning in chisq.test(tbl): Chi-squared approximation may be
incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

Chi-Squared Tests of Independence (2)

Exact Fisher test

```
fisher.test(tbl)


##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.4138
## alternative hypothesis: two.sided
```

Solution to Agresti exercise (1)

```
SeatBelt <- c("No","No","Yes","Yes")
Accident <- c("Fatal","Nonfatal","Fatal","Nonfatal")
Count <- c(54 , 10325, 25 , 51790)
Accdf <- data.frame(SeatBelt,Accident,Count)
Accdf
```

```
##   SeatBelt Accident Count
## 1       No    Fatal    54
## 2       No Nonfatal 10325
## 3      Yes    Fatal    25
## 4      Yes Nonfatal 51790
```

Solution to Agresti exercise (2)

```
rm(SeatBelt,Accident,Count)
names(dimnames(Accdf)) <- c("SeatBelt","Accident")
tab1 <- tapply(Accdf$Count,list(Accdf$SeatBelt,
Accdf$Accident),c);tab1

##     Fatal Nonfatal
## No     54    10325
## Yes    25    51790
```

Solution to Agresti exercise (3)

```
R.test <- prop.test(tab1)
odds <- R.test$estimate/(1-R.test$estimate)
names(odds) <- c("Odds1: F","Odds2: NF") ; odds

##     Odds1: F    Odds2: NF
## 0.0052300242 0.0004827187

OR <- odds[1]/odds[2]
names(OR) <- c("OR") ; OR

##       OR
## 10.83452
```

Solution to Agresti exercise (4)

```
theta <- odds[1]/odds[2]
ASE <- sqrt(sum(1/tab1))
# ASE
ASE
```

```
## [1] 0.242146
```

```
logtheta.CI <- log(theta) + c(-1,1)*1.96*ASE
# IC log(theta)
logtheta.CI
```

```
## [1] 1.908131 2.857343
```

```
# IC(OR)
exp(logtheta.CI)
```

```
## [1]  6.740479 17.415198
```

Références

Agresti, A. (2013). *Categorical Data Analysis*. Wiley edition.

Cornillon, P., Guyader, A., Husson, F., JÃ©gou, N., Josse, J., Kloareg, M., Matzer-Lober, E., and RouviÃ¨re, L. (2008). *Statistiques avec R*. PUR edition.

Greene, W. (2008). *Econometric Analysis, 6th*. Prentice-HallOxford : Clarendon Press edition.

Gujarati, D. (2003). *Basic Econometrics*. McGraw Hill edition.

Hensher, D. and Greene, W. (2003). The mixed logit model : the state of practice. *Transportation*, 30(2) :133–176.

Hensher, D., Rose, J., and Greene, W. (2005). *Applied Choice Analysis : A Primer*. New York : Cambridge University Press. edition.

McFadden, D. (2014). The new science of pleasure : consumer choice behavior and the measurement of well-being. In Hess, S. and Daly, A., editors, *Handbook of Choice Modelling*, pages 7–48. Edward Elgar, UK.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of of Applied Econometrics*, 64 :207–240.

Millot, G. (2012). *Comprendre et rÃ©aliser les tests statistiques Ã l'aide de R*. DeBoeck edition.

Munizaga, M. and Alvarez-Daziano, R. (2005). Testing mixed logit and probit models by simulation. *Transportation Research Record : Journal of the Transportation Research Board*, 1921 :53–62.