How to measure local development

A.A. 2013-2014

Dott. Giuliana Cortese
Department of Statistical Science
Padua University

Lecture 0

Sketch of the course

Course-Exam

- **Course (*28 hours*)**
  - Lessons (slides*)
  - Exercises (slides/blackboard)
  - Classroom Exercises
  - Homeworks
- **Exams**
  - written exam (with closed and open questions)
    - Computation of suitable statistical indicators (tendency and variation indices, correlation indicators)
    - Exploratory data analysis of data arranged in both graphs and tables
    - Evaluation of "significant" results

* revised from materials of Prof. Capizzi

Contacts

- *weekly office hours*:
  Thursday 14.00-16.00
  Department of Statistical Sciences
  via Cesare Battisti 241/243
- *Before the lessons*: Thursday 10.00 (same room)
- *Email*: giuliana.cortese@unipd.it
- *Tel*: 049-8274124

Books/References

- *Moore, D. S. The Basic practice of statistics, Freeman and Company, 1995 (Library Faculty of Statistics, Padua).*
- *Berenson, M., Levine M.L., Basic business statistics : concepts and applications – 7th Edion, Prentice Hall, 1999. (Library Dipartimento Marco Fanno, Padua);*
- Brase, C.H., Brase P.C., Understanding Basic Statistics [Paperback], 5th Edition, Brooks Cole, 2008.
- *Brase, C. H., Brase P. C., Understanding Basic Statistics, 6th Edition, Brooks Cole 2006.*
- Levine D.M., Krehbiel T.C. Berenson M.L., Business Statistics: A first course International version, 5th Edition, Pearson Higher Education, 2010
- Berenson M.L., Levine D.M., Krehbiel T.C., Basic Business Statistics, 11th Edition, Prentice Hall, 2009

Topics/Course Programme

- *Cases- qualitative and quantitative variables*
- *Frequency Tables; Graphs*
- *Measures of center and variation*
- *Random Variables* (expected values, variance, probability distributions)
  - Binomial and Normal distributions
- *Population and sample*
  - Sampling variation and sampling distribution
  - Central limit theorem
  - p-value
- *Confidence interval for means and proportions* (one-sample)
- *Hypothesis test for means and proportions* (one-sample)
- *Correlation and regression* (one-predictor)

Statistical Ingredients

- Cases (individuals, hospitals, countries, households, ecc…..)
- Variables
  - Income
  - Gender
  - Working Status
  - Height
  - Weight
  - Fertility rate

Statistical questions

1. What is “typical”?
2. How much “variety” is there?
3. How “certain” are we?
4. What should we compare this to?

What is "typical"?

- What's proportion of cases?
- What's average?
- How many cases?
- Where, when in particular?

Example

- Often used in journalism and politics
  - Restaurant examples
    - People are trying to sue McDonald's for making consumers fat→

      chain restaurants must be protected from frivolous lawsuits
  - Medical examples
    - Physician in Las Vegas closed his obstetric practice because of high malpractice premiums → Malpractice verdicts and insurance rates must be capped

How much "variety"?

- How extreme?
- How often?

How "certain"?

- How large is the margin error in the estimate?
- Is there a "significant" result/difference

What is "typical"?

Get beyond the single case to evaluate whether there's a broader problem.

- What proportion? What's the average?
- Restaurant examples
  - What proportion of consumer tort cases involve obesity?
  - What proportion of those cases go to trial?
  - What's the average verdict?
- Medical examples
  - What's the average malpractice insurance premium?
  - What proportion of medical practices close each year?
    - What proportion close because of insurance?

What is "typical"?
Response to "proof by average"

- The average malpractice premium is about $10K after taxes
- Only .05% of state lawsuits end in punitive damages

How "certain" are we?

Election results

- 46% of US voters are leaning toward Kerry
  - "Margin of error" +/- 3%
  - We're certain about the 1000 voters we talked to, but not about the others
- There's "no significant difference" between % voting for Kerry and Bush
  - There's a difference, but we're not sure which direction

How much "variety" is there?

- Get beyond the average
  - How extreme can it get?
  - How often does it get that extreme?
- The average malpractice premium is about $10K after taxes, but
  - it's twice that in obstetrics and surgical specialties
  - it's higher in certain counties
- Only .05% of state lawsuits end in punitive damages
  - but when damages are awarded they average over $1 million

Lecture 1

Data sets:
Cases and variables

Overview

- Data set
  - Cases
  - Variables
    - Interval (quantitative)
    - Nominal (qualitative)
      - Dichotomies and dummy coding
    - Ordinal (rank)
  - Discrete vs. Continuous variables

What is a data set?

- Organized information about a bunch of
  - People
  - Countries
  - ……
- Typically organized into
  - Cases (or observations)
  - Variables

Cases (or observations)

- Cases are the (horizontal) *rows* in a data set

Variables

- Variables are the (vertical) *columns*
- For each case, the variable has a particular *value*

Name	Gender	Graduating?	Class	Major	Age	Job hours	Children?
Theda Skocpol	F	no	sr	criminology	21	10	no
Jane Addams	F	yes	jr	sociology	23	15	no
Andrew Greeley	M	no	sr	criminology	23	25	yes
Karl Marx	M	no	sr	criminology	24	35	no
Georg Simmel	M	no	sr	criminology	21	34-40	no

In this data set
• each case is a person
• one variable is Gender (its possible values are F and M).

Country	Working women	GDP per person	Urban	Religion
France	44%	$19,510	73%	Catholic
Britain	46%	$17,160	89%	Protestant
W. Germany	39%	$14,730	86%	Protestant
Italy	30%	$18,090	67%	Catholic
Netherlands	31%	$17,780	89%	Protestant
Spain	22%	$13,400	76%	Catholic
Ireland	31%	$12,830	57%	Catholic

In this data set
- each case is a country
- one variable is (name of) Country (its values are France, Britain...)

Types of variables (levels of measurement)

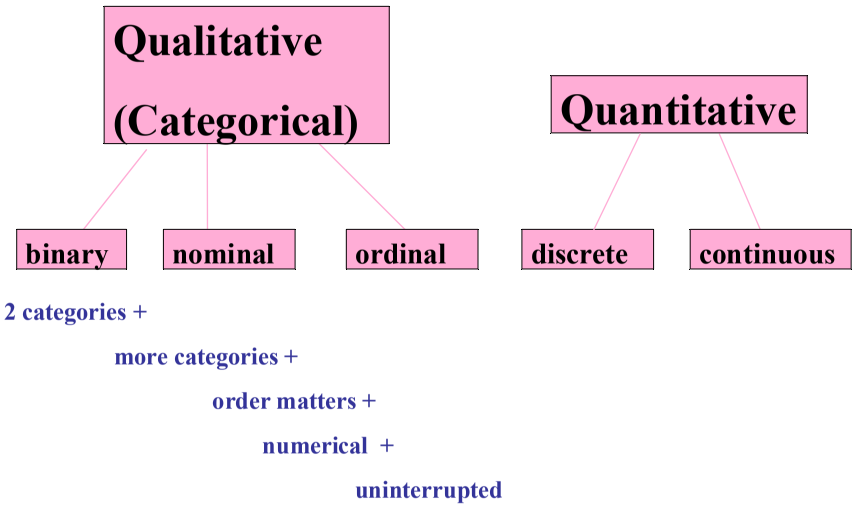
- **Qualitative** variables in which there is no measuring involved (favorite color, religion, city of birth, favorite sport, etc.)
  - "nominal"
  - "ordinal"
- **Quantitative** variables measured on a **numeric** scale (Height, weight, response time, subjective rating of pain, temperature, and score on an exam, etc.)
  - "interval"
  - "ratio"

CAUTION: the type of variable you have determines the type of statistics and analysis you can do.

Rank	Team	Computer Avg.	Schedule Strength	Schedule Rank	Losses	Quality Wins	Total
1	Miami (Fla.)	1.17	19	0.76	0	0	2.93
2	Ohio State	1.67	20	0.8	0	-0.5	3.97
3	Georgia	3.17	5	0.2	1	0	8.37
4	USC	3.67	1	0.04	2	-0.2	10.51
5	Iowa	4.83	49	1.96	1	0	10.79
6	Washington St.	7	21	0.84	2	-0.7	16.14
7	Oklahoma	6.33	14	0.56	2	-0.1	16.79
8	Kansas State	10.67	54	2.16	2	-0.7	20.13
9	Notre Dame	6.83	15	0.6	2	0	20.93
10	Texas	9.5	22	0.88	2	-0.3	21.08
11	Michigan	9.33	2	0.08	3	0	23.91
12	Penn State	13.33	16	0.64	3	0	26.97
13	Colorado	15.17	10	0.4	4	-0.3	33.27
14	Florida State	13.83	3	0.12	4	0	33.95
15	West Virginia	17.33	41	1.64	3	0	35.97

In this data set
- each case is a football team
- one variable is rank, with values 1,2,3,…

Types of Variables: Overview



Categorical Variables

- Nominal ("qualitative") variables ←→named categories
- Why is it called nominal
  - Latin *nomen* = name
  - The values are just names
- Order doesn't matter!
- Values are different, but not more or less
  - Phone numbers
  - Jobs: butcher, baker, candlestick maker

Dummy coding

- Take a dichotomy
- Call one value 1, the other 0
  - E.g., Male=1, Female=0
  - Or Female=1, Male=0
  - It doesn't matter which, as long as you remember

- Example: dichotomies coded as dummies

Name	Male (1 if yes)	Graduating (1 if yes)	Class	Major	Age	Job hours	Children (1 if yes)
Theda Skocpol	0	0	sr	criminology	21	10	0
Jane Addams	0	1	jr	sociology	23	15	0
Andrew Greeley	1	0	sr	criminology	23	25	1
Karl Marx	1	0	sr	criminology	24	35	0
Georg Simmel	1	0	sr	criminology	21	34-40	0

Categorical Variables

- Binary (Dichotomous) : nominal variable with only two possible values
  - Status
    - Dead/alive
    - Disease/no disease
  - Experimental status
    - Treatment/placebo
    - Exposed/Unexposed
    - Heads/Tails
  - Gender
    - Male/Female

Categorical Variables

- Nominal Variables (more than 2 categories)
  - Treatment groups
  - Exposure groups
  - Working status
  - The blood type of a patient (O, A, B, AB)
  - Marital status
  - Occupation

Categorical Variables

- Ordinal variable – Ordered categories. Order matters!
  - Staging in breast cancer as I, II, III, or IV
  - Birth order—1st, 2nd, 3rd, etc.
  - Letter grades (A, B, C, D, F)
  - Ratings on a scale from 1-5
  - Ratings on: always; usually; many times; once in a while; almost never; never
  - Age in categories (10-20, 20-30, etc.)
  - Shock index categories (Kline et al.)

cases	agegp	alcgp	tobgp
1	25-34	0-39g/day	0-9g/day
2	25-34	0-39g/day	10-19g/day
3	25-34	40-79g/day	20-29
4	25-34	80-119	30+
5	25-34	80-119	10-19g/day
6	55-64	120+	20-29
7	65-74	0-39g/day	20-29
8	65-74	40-79g/day	0-9g/day
9	75+	120+	30+

Example

Data from a case-control study of (o)esophageal cancer in Ille-et-Vilaine, France

- "agegp" Age group

code	categories
1	25--34 years
2	35--44
3	45--54
4	55--64
5	65--74
6	75+

- "alcgp" Alcohol consumption

code	categories
1	0--39 gm/day
2	40--79
3	80--119
4	120+

- "tobgp" Tobacco consumption

code	categories
1	0-- 9 gm/day
2	10—19
3	20--29
4	30+

Ordinal (rank) variables

- *Rank order* of values
- *intervals* between values are not meaningful/comparable
- Examples:
  - In a horserace
    - time is interval
    - but place is ordinal

  - Military rank
  - Lower/middle/upper class
  - Disagree strongly—disagree—agree—agree strongly

Quantitative variables

- Discrete
- Continuous

Quantitative Variables

- Discrete Numbers – a limited set of distinct values, such as whole numbers.
  - Number of new AIDS cases in CA in a year (counts)
  - Years of school completed
  - The number of children in the family (cannot have a half a child!)
  - The number of deaths in a defined time period (cannot have a partial death!)
  - Roll of a die

Quantitative Variables

- Continuous Variables Can take any number within a defined range and may be arithmetically manipulated.
  - Time
  - Age
  - Height
  - Time-to-event (survival time)
  - Age
  - Blood pressure
  - Serum insulin
  - Speed of a car
  - Income
  - Respiratory rate

Quantitative ("interval")

- How much more?
- Why is it called interval?
  - You can quantify distance between cases
  - You can talk about differences, the zero is arbitrary (Ex. Temperature, date)

Quantitative ("ratio")

- How many times more?
- Why is it called interval?
  - You can also take ratios between cases
  - The zero is meaningful (Ex. Weight, Respiratory rate, age)

Examples

Name	Gender	Graduating?	Class	Major	Age	Job hours	Children?
Theda Skocpol	F	no	sr	criminology	21	10	no
Jane Addams	F	yes	jr	sociology	23	15	no
Andrew Greeley	M	no	sr	criminology	23	25	yes
Karl Marx	M	no	sr	criminology	24	35	no
Georg Simmel	M	no	sr	criminology	21	34-40	no

E.g., Karl works 35 hours, Theda works 10
Interval = 35-10 = 25 hours
Karl works 25 more hours

Example

- Rankings of football teams
  - Ordinal variables in black; continuous variables in gray

Rank	Team	Computer Avg.	Schedule Strength	Schedule Rank	Losses	Quality Wins	Total
1	Miami (Fla.)	1.17	19	0.76	0	0	2.93
2	Ohio State	1.67	20	0.8	0	-0.5	3.97
3	Georgia	3.17	5	0.2	1	0	8.37
4	USC	3.67	1	0.04	2	-0.2	10.51
5	Iowa	4.83	49	1.96	1	0	10.79
6	Washington St.	7	21	0.84	2	-0.7	16.14
7	Oklahoma	6.33	14	0.56	2	-0.1	16.79
8	Kansas State	10.67	54	2.16	2	-0.7	20.13
9	Notre Dame	6.83	15	0.6	2	0	20.93
10	Texas	9.5	22	0.88	2	-0.3	21.08
11	Michigan	9.33	2	0.08	3	0	23.91
12	Penn State	13.33	16	0.64	3	0	26.97
13	Colorado	15.17	10	0.4	4	-0.3	33.27
14	Florida State	13.83	3	0.12	4	0	33.95
15	West Virginia	17.33	41	1.64	3	0	35.97

Caution

- Not all numbers are interval variables
  - E.g., phone numbers
    - My office phone is 6883768
    - One of Shelley's is 2922115
    - Do I have *more* phone number?
      - 6883768- 2922115=3961653
    - No! Our numbers are just different!

Summary of variable types

Type of variable	Order	Distance	All Relationship
Nominal			
Ordinal	x		
Interval	x	x	
Ratio	x	x	x

Exercise

- Which are the interval variables? Nominal? Dichotomies?

Country	Working women	GDP per person	Urban	Religion
France	44%	$19,510	73%	Catholic
Britain	46%	$17,160	89%	Protestant
W. Germany	39%	$14,730	86%	Protestant
Italy	30%	$18,090	67%	Catholic
Netherlands	31%	$17,780	89%	Protestant
Spain	22%	$13,400	76%	Catholic
Ireland	31%	$12,830	57%	Catholic

Exercise

- Find dichotomies and code as dummies

Country	Working women	GDP per person	Urban	Religion
France	44%	$19,510	73%	Catholic
Britain	46%	$17,160	89%	Protestant
W. Germany	39%	$14,730	86%	Protestant
Italy	30%	$18,090	67%	Catholic
Netherlands	31%	$17,780	89%	Protestant
Spain	22%	$13,400	76%	Catholic
Ireland	31%	$12,830	57%	Catholic

Example/exercise

- Find the dichotomies

Name	Gender	Graduating?	Class	Major	Age	Job hours	Children?
Theda Skocpol	F	no	sr	criminology	21	10	no
Jane Addams	F	yes	jr	sociology	23	15	no
Andrew Greeley	M	no	sr	criminology	23	25	yes
Karl Marx	M	no	sr	criminology	24	35	no
Georg Simmel	M	no	sr	criminology	21	34-40	no

- Treat football rankings as interval
- Treat Disagree strongly—disagree—agree—agree strongly
  - As nominal. Or code as
    - 0—1—2—3
  - and treat as interval