Social Studies and Communication

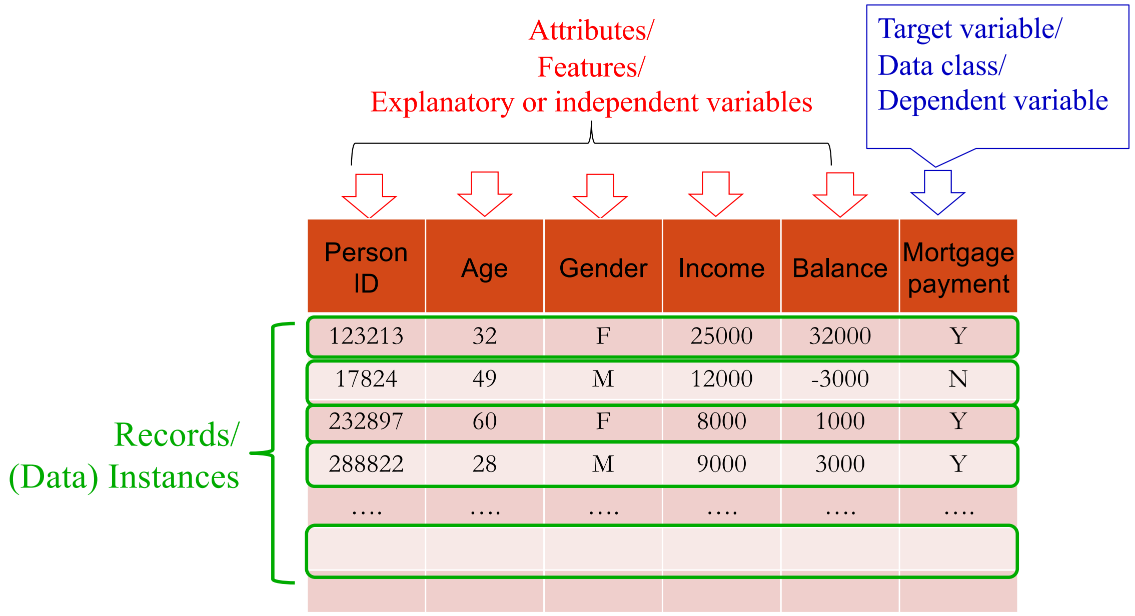
Introduction to Strategic Data Science

2nd lecture

Ryuichiro Ishikawa

Online Lecture: 4th on Mon

Individual data



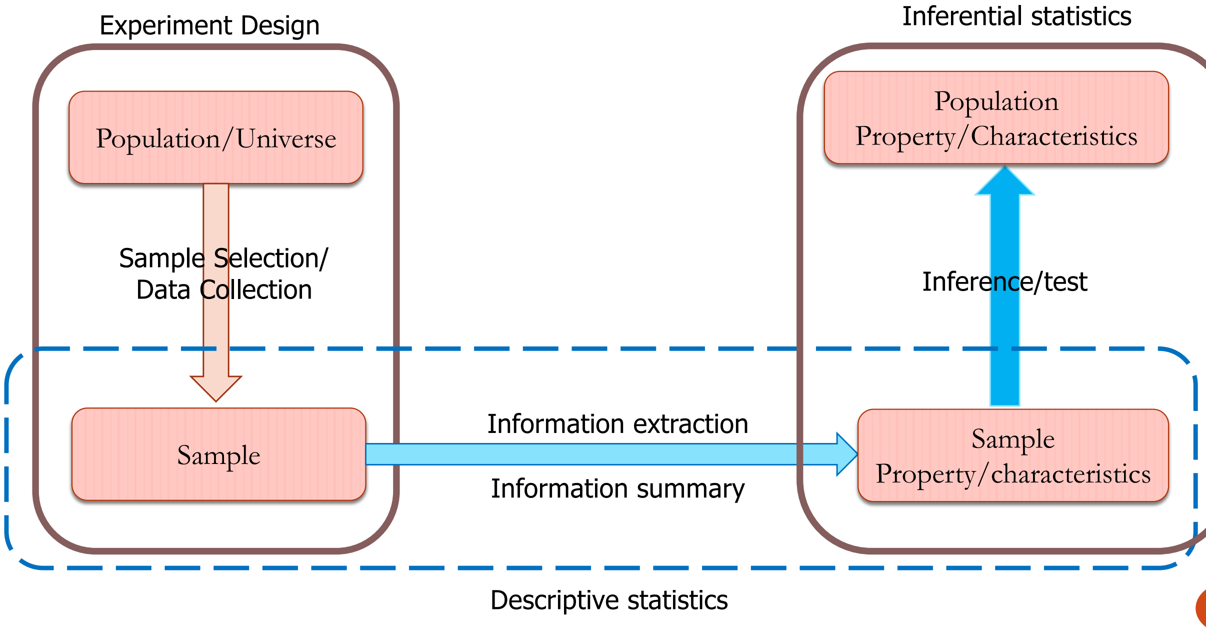
Attributes/
Features/
Explanatory or independent variables

Target variable/
Data class/
Dependent variable

Records/
(Data) Instances

Person ID	Age	Gender	Income	Balance	Mortgage payment
123213	32	F	25000	32000	Y
17824	49	M	12000	-3000	N
232897	60	F	8000	1000	Y
288822	28	M	9000	3000	Y
….	….	….	….	….	….

Structure of experiments & statistics



Experiment Design

Population/Universe

Sample Selection/ Data Collection

Sample

Information extraction

Information summary

Inferential statistics

Population Property/Characteristics

Inference/test

Sample Property/characteristics

Descriptive statistics

Stop Churning!!

- A cell phone company has a problem with customer retention!
  - Communications companies are now engaged in battles to attract each other's customers while retaining their own.
  - Customers switching from one company to another is called “**churn**”.
  - Our task is to devise a precise, step by step plan for how the data science team should use the vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

Data of Churning

- We have the following data; 02churn.csv:
  - The ‘csv’ extension is a data file of text type, separated by commas. It is a common to exchange data.
  - You can open any text editor app or Excel.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	OVER_15MINS_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE
sample_01	zero	31953	0	6	313378	161	0	4	unsat	little	STAY
sample_02	one	36147	0	13	800586	244	0	6	unsat	little	STAY
sample_03	one	27273	230	0	305049	201	16	15	unsat	very_little	STAY
sample_04	zero	120070	38	33	788235	780	3	2	unsat	very_high	LEAVE
sample_05	one	29215	208	85	224784	241	21	1	very_unsat	little	STAY
sample_06	zero	133728	64	48	632969	626	3	2	unsat	high	STAY
sample_07	zero	42052	224	0	697949	191	10	5	very_unsat	little	STAY
sample_08	one	84744	0	20	688098	357	0	5	very_unsat	little	STAY

The meaning of attributes

Variable	Explanation
COLLEGE	Is the customer college educated? (zero: No; one: Yes)
INCOME	Annual Income ($US)
OVERAGE	Average overcharged per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
OVER_15MINS_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self reported usage level
LEAVE	Did the customer stayor leave (churn)?

Kind of data

- Data is collected in the different styles.
  - We can classify the data types as follows:

Qualitative data	**Nominal scale**	Nominal Quantification.	College
	Ordinal scale	In addition to nominal scale, the order also matters.	Say, 1 as prefer, 5 as not preferable
Quantitative data	**Interval scale**	In addition to ordinal scale, the number intervals also matters.	Time, temperature, and so on.
	Proportional scale	In addition to Interval scales, the ratio also matters.	Income, House and so on.

Descriptive statistics

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	OVER_15MINS_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE
sample_01	zero	31953	0	6	313378	161	0	4	unsat	little	STAY
sample_02	one	36147	0	13	800586	244	0	6	unsat	little	STAY
sample_03	one	27273	230	0	305049	201	16	15	unsat	very_little	STAY
sample_04	zero	120070	38	33	788235	780	3	2	unsat	very_high	LEAVE
sample_05	one	29215	208	85	224784	241	21	1	very_unsat	little	STAY
sample_06	zero	133728	64	48	632969	626	3	2	unsat	high	STAY
sample_07	zero	42052	224	0	697949	191	10	5	very_unsat	little	STAY
sample_08	one	84744	0	20	688098	357	0	5	very_unsat	little	STAY

- (Sample) average $\bar{X} = \sum_i x_i / N$ where x_i is sample i's data of attribute x;

- Sample variance: $s^2 = \sum_i \frac{(x_i - \bar{X})^2}{N-1}$.