Data Analysis and Hypothesis Testing Using the Python ecosystem

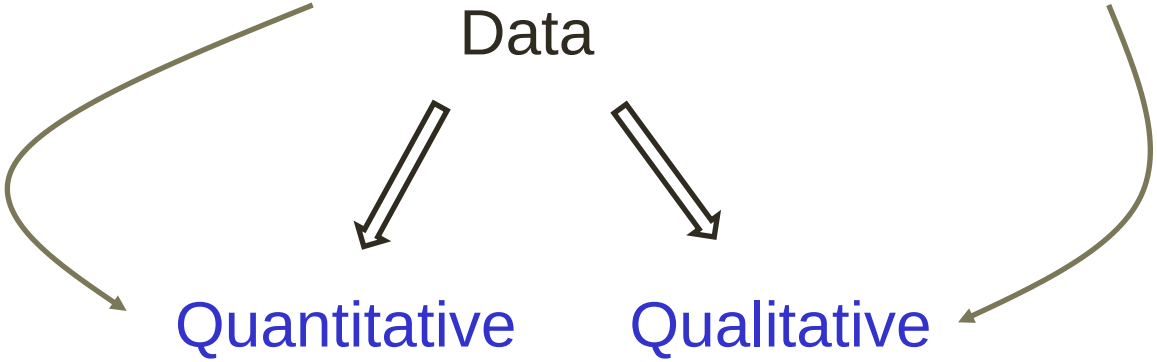
An introduction to the quantitative research paradigm

Stavros Demetriadis

sdemetri@csd.auth.gr
http://mlab.csd.auth.gr/sdemetri

Data

- **Data** are abstractions that **reveal perspectives** of the world we live in
- Usually available as **collections of values** or **networks of concepts**



- A **value** is an expression which cannot be evaluated any further (Wikipedia)
  - 3 is a value, 1 +2 is *not* a value
- A **concept** is an abstraction useful for categorization of world entities
  - A **semantic network** (conceptual network) represents semantic relations between concepts (Wikipedia)

Quantities

- **Quantitative** data are produced by *measurement*: comparison to a given measuring instrument
  - For example: learners' performance in a standardized test

Table 5 The students' scores in the pre- and post-test questionnaires

		Pre-test		Post-test	
	n	M	SD	M	SD
Control	32	10.94	5.41	10.13	4.48
U Treatment	32	10.65	3.84	11.95	3.99
D Treatment	32	10.45	4.06	14.01	3.68

Tables from: Tegos, S., Demetriadis, S., Papadopoulos, P., & Weinberger, A. (2016). Conversational Agents for Academically Productive Talk: A Comparison of Directed and Undirected Agent Interventions. *International Journal of Computer Supported Collaborative Learning* (to appear)

Qualities

- **Qualitative** data produced by analysis of descriptions
  - For example: analysis of students' discourse
  - Qualitative data available through **depictive code** (for example, images, videos) are also transcribed as descriptions

Table 3 A dialogue excerpt showcasing a directed agent intervention

	User	Message
1.	Philip:	That doesn't seem right. The menu closes instantly if you move your mouse pointer out of the popup.
2.	Jason:	That's correct.
3.	Philip:	OK
4.	Agent:	*Jason, do you believe the Accot-Zhai law somehow relates to mouse movement in menu?*
5.	Jason:	Hmm, please give me a minute to respond.
6.	Philip:	ok no problem, ask me if you need anything
7.	Jason:	Do you remember the mathematical expression?
8.	Philip:	T=a*b*(D/S)☺
9.	Jason:	*[Submitted Answer]* The Accot-Zhai (or steering) law, which predicts the time required to steer a pointing device through a 2D tunnel (T=a*b*(D/S)), relates to the top cascading menu as it does not provide users with the necessary time to navigate through the hierarchical menu options without closing.
10.	Philip:	That's true because, although the menu appears to be OK in size, it could be improved by expanding the 'active' pointer region or placing menu items closer to each other
11.	Jason:	That's a nice suggestion actually, we could also add a delay so that the sub-menu does not close immediately while moving the pointer between menu items
12.	Philip:	Let's move on

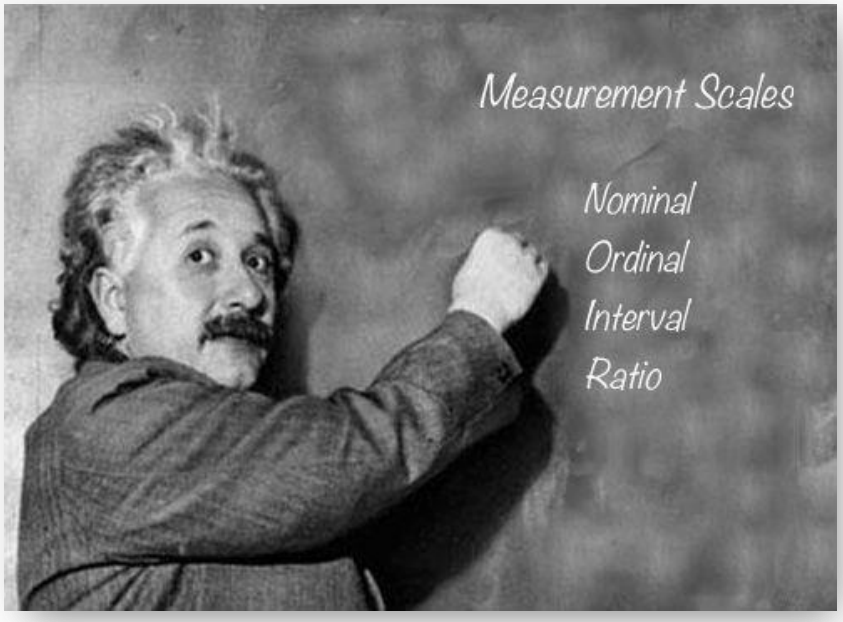
The two ‘worlds’ interact

- **Qualitative decisions** are important in the quantitative world
  - For example: how to develop and validate a measuring instrument?
- **Quantitative processing** is important in the qualitative world
  - Frequencies → processing
  - Scheme-based classification → processing
  - ……
- **Mixed methods research**
  - the mixing of qualitative and quantitative data and methodologies/paradigms in a research study

Measuring

- A **measure** (variable): what do we measure?
  - For example: learner's learning performance
- A **measuring instrument**: how do you measure the variable?
  - For example: with a standardized knowledge test
- But not all measurements are the same

Levels of measurement 1/2



- **Nominal** (categorical)
  - Data are classified in categories with **no particular order**: e.g. boys and girls

- **Ordinal**
  - Data are ordered but **distances** between measurement **has no meaning**
  - For example: a Likert scale 1 ('Strongly Disagree') to 5 ('Strongly Agree')
  - 5 ('Strongly Agree') is 'more' than 4 ('Agree') but the distance between 5 and 4 is meaningless

  - The mean of an ordinally-measured variable is a meaningful statistic BUT prefer reporting **mode or median** (not mean) for central tendency

Levels of measurement 2/2

- **Interval**
  - **Distance** between data is **meaningful** but not the ratio (the scale has no absolute zero)
  - For example: when referring to temperature measurements 'distances' (e.g. 5^{o} to 10^{o}) are meaningful. But stating that '20^{o} is double as hot as 10^{o}' is meaningless.
- **Ratio**
  - In ratio level of measurement **ratios** and an **absolute zero are meaningful**.
  - For example: measuring the learners' performance in a scale of 0-10 scoring 0 is meaningful ('nothing performed'). Also, scoring 10 is performing twice as good as scoring 5.
  - Ratio scales is what we need to apply **meaningful statistical analysis.**
  - For example central tendency (mode, median, or mean), standard deviation,…

Research Design in Social/Life Sciences

- *Depending on Sampling*:
- **Random** assignment ➔ Randomized experimental design
- **Non-random** assignment ➔ Quasi-experimental design
  (for example, groups taken intact)

- *Depending on Groups & Pre/Post Test*:
- Post-test only:

R	X	O
R		O

- Pre/Post Test

R	O	X	O
R	O		O

more@pytolearn

Key issues when measuring

- **Reliability**: how reliable are the measurements?
- **Validity**: are the measuring instrument(s) valid?
- **Generalizability**: after analyzing data can conclusions be generalized?

more@pytolearn

Reliability

- **Reliability** in statistics and psychometrics is the **overall consistency** of a measure (Wikipedia)

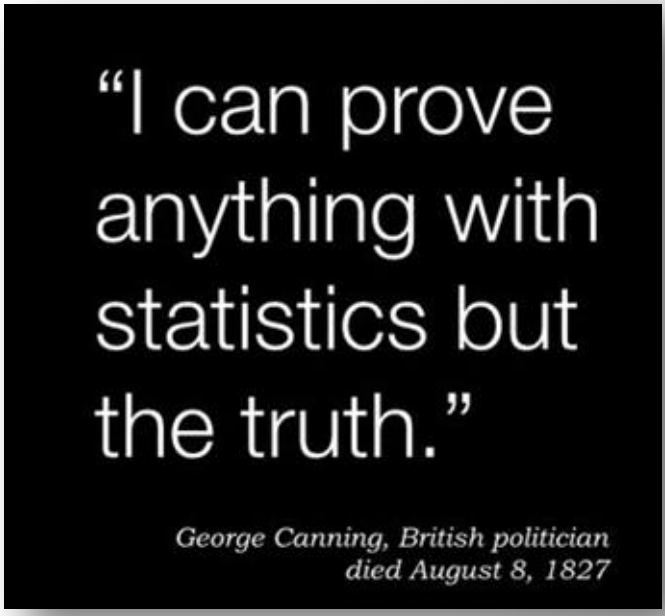
- In other words: reliability is the quality of ensuring that under ***similar*** conditions the instrument will produce ***similar*** measurements – thus, results are *repeatable*
- Various types of reliability: inter-rater, test-retest, etc.

- Common reliability measure: Cronbach's alpha
  - Measure of internal consistency, that is, how closely related a set of items are as a group (SPSS FAQ, Univ. of Virginia)
  - Acceptable: **0.8 > a >= 0.7**

Validity

- **Validity** is the extent to which a concept, conclusion or measurement is well-founded and corresponds accurately to the real world.
- Does the tool **measure what it claims to measure?** (Wikipedia)

- Many dimensions of validity:
  - Construct validity
  - Internal validity
  - External validity
  - ……



Is this statement valid?

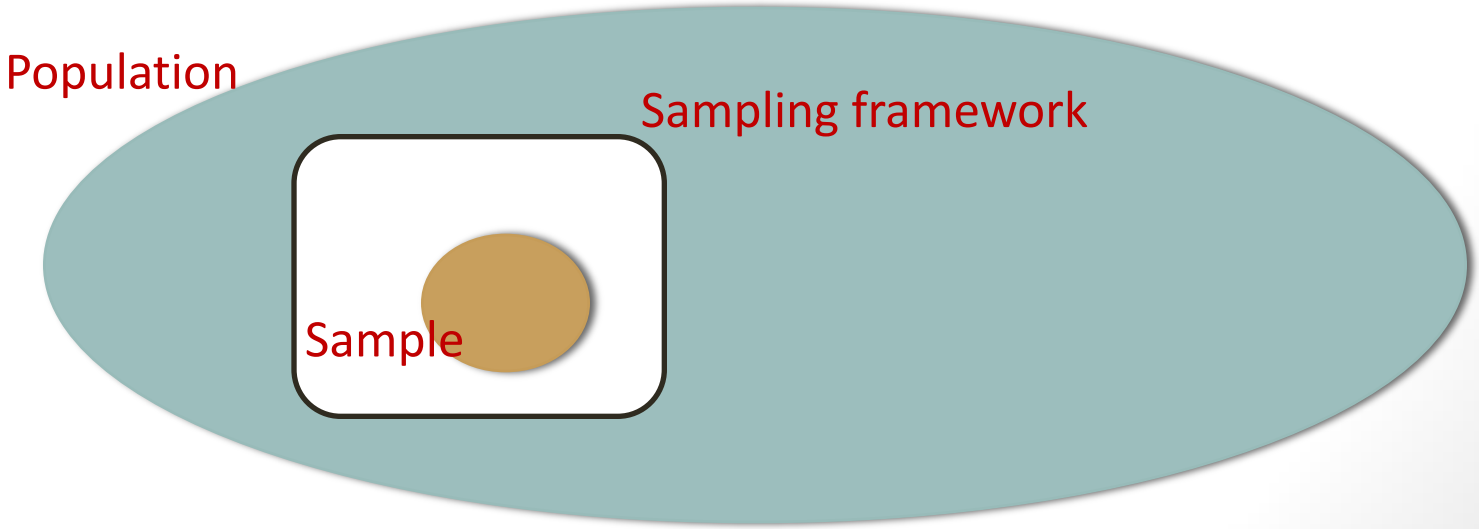
Reliability and validity are not the same

- …But they are both indicators of quality research

Source

Generalizability

- The extension of research findings and conclusions from a study conducted on a **sample population to the population** at large (Colorado State University)
- In other words: what we find in a sample is valid for the whole population?



True score theory

True score

Measurement

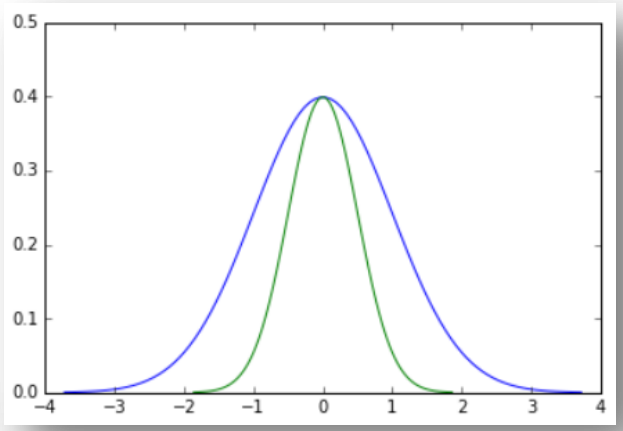
$$x = T + e$$

Error

Random Error → $e_r + e_s$ ← **Systematic** Error

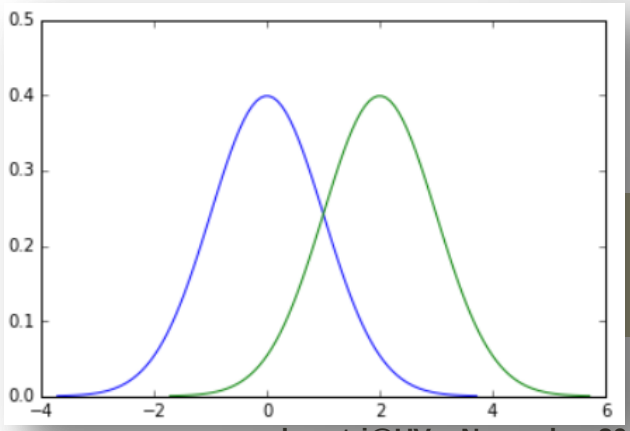
(affects variability ➔ 'noise')

(affects mean ➔ 'bias')



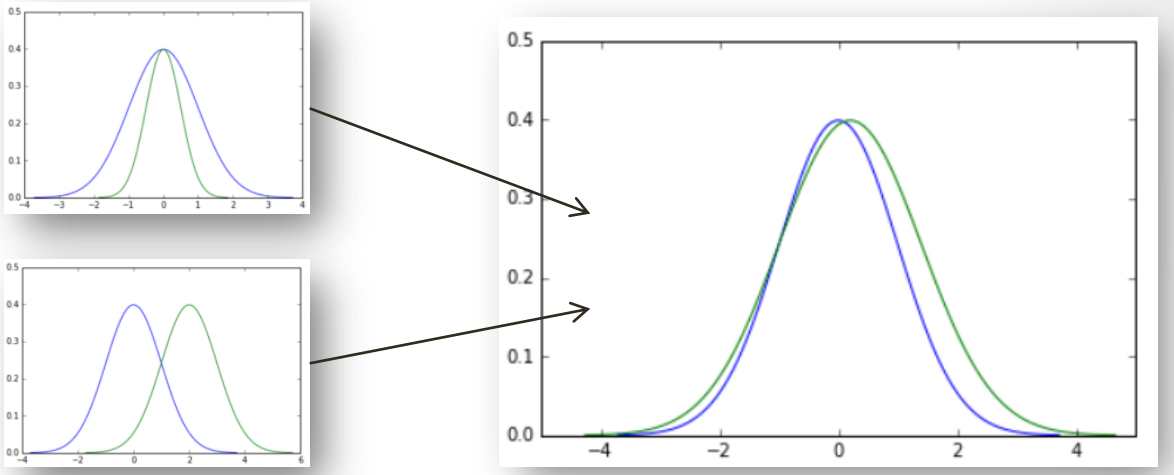
x

T



High quality research features:

- **High Reliability**: by eliminating mainly *systematic error*



- **High Validity**: through *argumentation* or comparison with other *validated data sets*
- **Representative sampling**: eliminating *sampling error* (by increasing sample size and considering stratified sampling)
  - 'Stratified': sampling according to subpopulations

I got my data, now what?

- You need a tool to **bring** your data in the computer and **represent** them in a meaningful way

	Country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Code											
ESP	Spain	44397319	45226803	45954106	46362946	46576897	46742697	46773055	46620045	46480882	46418269
FRA	France	63621376	64016229	64374990	64707044	65027512	65342776	65659790	65972097	66495940	66808385
GRC	Greece	11020362	11048473	11077841	11107017	11121341	11104899	11045011	10965211	10892413	10823732
IRL	Ireland	4273591	4398942	4489544	4535375	4560155	4576794	4586897	4598294	4617225	4640703
ITA	Italy	58143979	58438310	58826731	59095365	59277417	59379449	59539717	60233948	60789140	60802085
MLT	Malta	405308	406724	409379	412477	414508	416268	419455	423374	427364	431333
PRT	Portugal	10522288	10542964	10558177	10568247	10573100	10557560	10514844	10457295	10401062	10348648
CYP	Cyprus	1048293	1063040	1077010	1090486	1103685	1116644	1129303	1141652	1153658	1165300

- **Data 'wrangling'** (or 'munging'): the process of manually **converting** or **mapping** data from one "raw" form into another format that allows for **more convenient consumption** of the data with the help of semi-automated tools (Wikipedia)

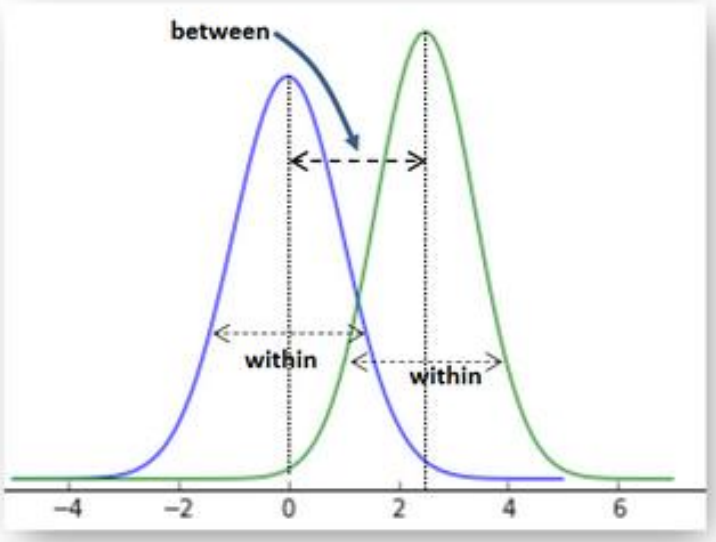
…and what is 'hypothesis testing'?

- A **hypothesis** *is a specific statement of prediction relevant to the phenomenon under study*.

- *Example*:
- *A research question*: Does background music in a multimedia learning environment have a positive/negative impact on students who use this environment to learn?
- *A null hypothesis* H_0: "**Background music has no impact whatsoever on students' learning**"
- Based on our data we either **reject** or '**fail to reject**' the null hypothesis - But how?

The rationale for hypothesis testing

$$\text{measure} = \frac{\text{signal}}{\text{noise}} = \frac{\textbf{Between} \text{ groups variability}}{\textbf{Within} \text{ groups variability}}$$



- If **between groups** variability is found to be **very large** compared to **within groups** then something beyond pure chance is happening

more@pytolearn

So, what exactly do we do?

Procedure	Example: t-test
Define a statistic	$t = \frac{M_1 - M_2}{s_p \sqrt{\frac{2}{n}}}$
Compute the value of the statistic based on experimental data	t = 3.706
Check the statistic distribution and **find** the probability that such a value appears	p = 0.0004
Compare to the threshold value 'a' (usually set to 0.05)	p < a (0.05)
Decide: 1) p <= a → significant 2) p > a → 'non significant'	**Statistically significant** → The two samples come from different populations → The treatment factor had an impact

Python ecosystem (PE) tools

- *Data management (wrangling or munging):* **pandas**
- *Statistics*: **Scipy, statsmodels**, …

- PE is a general-purpose programming environment (not a statistical package)

- **Pros**: you can implement and streamline any kind of data analysis, you can write your own data processing code

- **Cons**: if your focus is more specific, consider using:
  - R: language and environment for **statistical computing**
  - SPSS, SAS, etc.: **statistical packages**

- Comparison of statistical packages@wikipedia