Pere-Pau Vázquez

IDI – Quantitative and Qualitative Methods for Human-Subject Experiments

Outline

- Motivation
- Validity of experiments
- Experiment design
- Data analysis

Outline

- **Motivation**
- Validity of experiments
- Experiment design
- Data analysis

Motivation

- Measuring the response of humans to different experiments is the only way to:
  - Evaluate how humans perceive, manipulate, reason with applications or webpages
  - Measure utility of applications and webpages
- Key issue in software development
- It is important to do **before** launching any product!

Outline

- *Motivation*
- **Validity of experiments**
- Experiment design
- Data analysis

Validity of experiments

- Empirical method:
  - Develop a hypothesis, perhaps based on a theory
  - Make the hypothesis testable
  - Develop an empirical experiment
  - Collect and analyse data
  - Accept or refute the hypothesis
  - Relate the results back to the theory
  - If worthy, communicate the results to scientific community

Validity of experiments

- Experimental Validity
  - Does experiment really measure what we want it to measure?
  - Do our results really mean what we think (and hope) they mean?
  - Are our results reliable?
    - If we run the experiment again, will we get the same results?
    - Will others get the same results?

Validity of experiments

- Experimental variables:
  - Independent Variables
    - What the experiment is studying
    - Occur at different levels
      - Example: stereopsis, at the levels of stereo, mono
    - Systematically varied by experiment

Validity of experiments

- Experimental variables:
  - Dependent Variables
    - What the experiment measures
    - Assume dependent variables will be affected by independent variables
    - Must be measurable quantities
      - Time, task completion counts, error counts, survey answers, scores, etc.
      - Example: VR navigation performance, in total time

Validity of experiments

- Experimental variables:
  - Independent variables can vary in two ways
    - Between-subjects: each subject sees a different level of the variable
      - Example: 1/2 of subjects see stereo, 1/2 see mono
    - Within-subjects: each subject sees all levels of the variable
      - Example: each subject sees both stereo and mono

Validity of experiments

- Experimental variables:
  - Confounding factors (or confounding variables)
    - Factors that are not being studied, but will still affect experiment
      - Example: stereo condition less bright than mono condition
    - Important to predict and control confounding factors, or experimental validity will suffer
      - E. g.: Mono vs stereo and brightness

MOVING

Outline

- *Motivation*
- *Validity of experiments*
- **Experiment design**
- Data analysis

Experiment design

- To avoid skewing effects, experiments must be designed carefully
  - E. g.: Learning a technique
    - After N repetitions of the same experiment, the user will go fast to solve the same problem
  - E. g.: Suffering fatigue
    - After N repetitions, if the task requires physical effort, the performance may suffer

Experiment design

- Counterbalancing design:
  - Avoid learning/fatigue effects by randomizing the tasks
    - Randomizing does not necessarily mean random, but sorting adequately users and conditions (systematic variation)

Experiment design

- Let's imagine we have 10 subjects and we want to test solving the same task (e. g. buying a book) using two different websites:

Subjects	First shopping	Second shopping
1, 3, 5, 7, 9	Website A	Website B
2, 4, 6, 8, 10	Website B	Website A

Experiment design

- Let's imagine we want to test solving the same task (e. g. buying a book) using three different devices (desktop, tablet, and mobile).
  - We will have the following conditions:

	Website	
Device	**Website A**	**Website B**
Smartphone		
Tablet		
Desktop		

Experiment design

- Say that we want each user to perform each task 4 times
  - We will have 3 (devices) x 2 (websites) x 4 (repetitions) = 24 tasks
    - Note that this grows with a factorial explosion!!!
- To ensure reliability, those tests must be performed in the adequate order
  - Different for each subject

Experiment design

- Latin squares (Not what you find on Wikipedia!):
  - Tabular expression of systematic variations
  - Can be used to adequately sort experimental tasks
  - Counterbalances to avoid confounding factors
    - Within-subjects variables: control fatigue and learning effects
    - Between-subjects variables: control other factors that change with time (e. g. network speed, cache contents)

Experiment design

- Latin squares. Properties:
  - Every level appears in the every position the same number of times
  - Every level is followed by every other level
  - Every level is preceded by every other level

Experiment design

- Latin squares. Examples:

2x2

1	2
2	1

6x3

1	2	3
2	3	1
3	1	2

1	3	2
2	1	3
3	2	2

4x4

1	2	3	4
2	4	1	3
3	1	4	2
4	3	2	1

Experiment design

- Studying the previous example (3 devices) x 2 websites with 4 repetitions:
  - Form a Cartesian product of latin squares:
    - 6x3 (devices) x 2x2 (conditions)
    - This will counterbalance properly a group of 12 subjects

Experiment design

Subject	Presentation order
1	1A, 1B, 2A, 2B, 3A, 3B
2	1B, 1A, 2B, 2A, 3B, 3A
3	2A, 2B, 3A, 3B, 1A, 1B
4	2B, 2A, 3B, 3A, 1B, 1A
5	3A, 3B, 1A, 1B, 2A, 2B
6	3B, 3A, 1B, 1A, 2B, 2A
7	1A, 1B, 3A, 3B, 2A, 2B
8	1B, 1A, 3B, 3A, 2B, 2A
9	2A, 2B, 1A, 1B, 3A, 3B
10	2B, 2A, 1B, 1A, 3B, 3A
11	3A, 3B, 2A, 2B, 1A, 1B
12	3B, 3A, 2B, 2A, 1B, 1A

Experiment design

- More examples in the PDF.

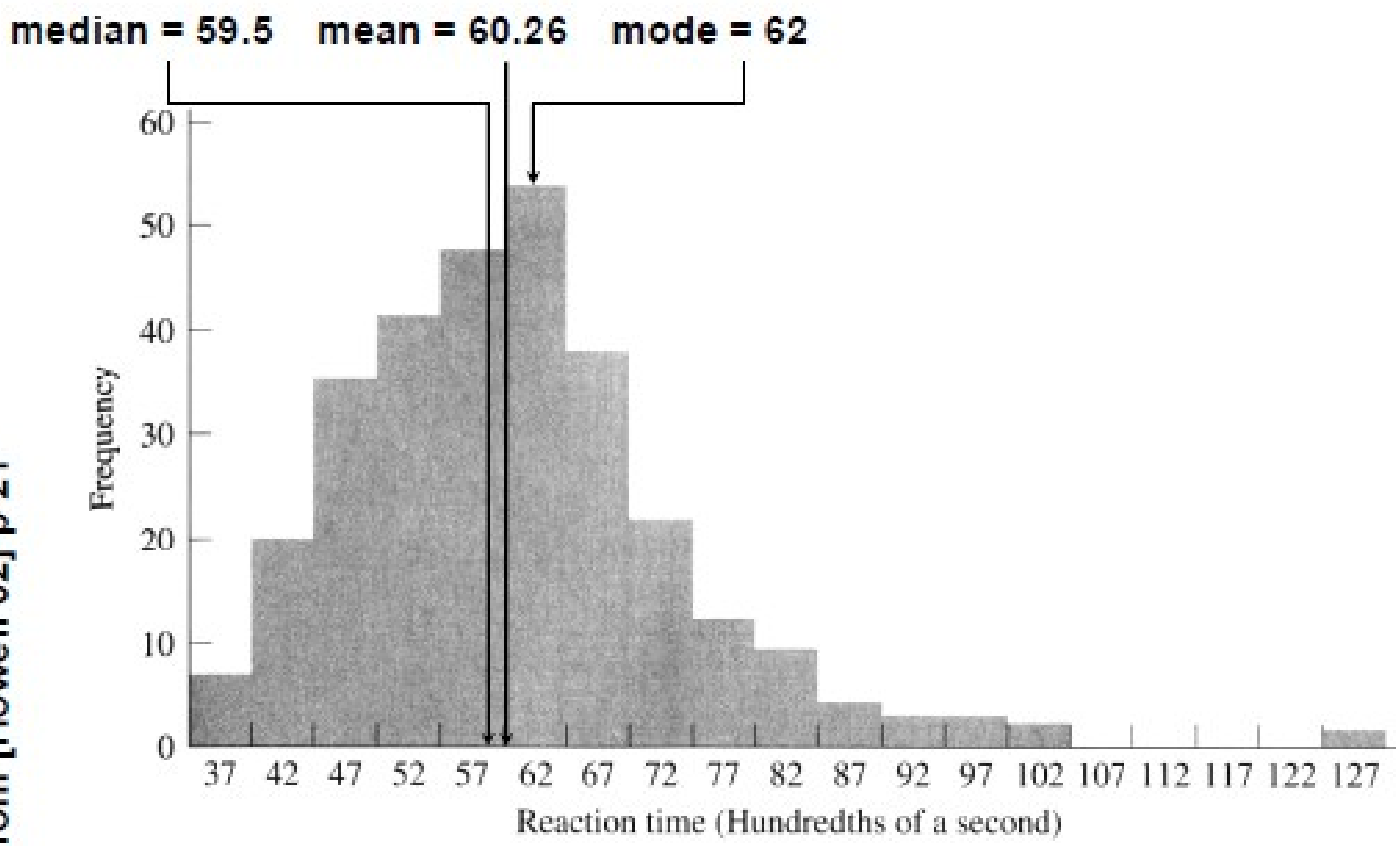
Outline

- *Motivation*
- *Validity of experiments*
- *Experiment design*
- **Data analysis**

Data analysis

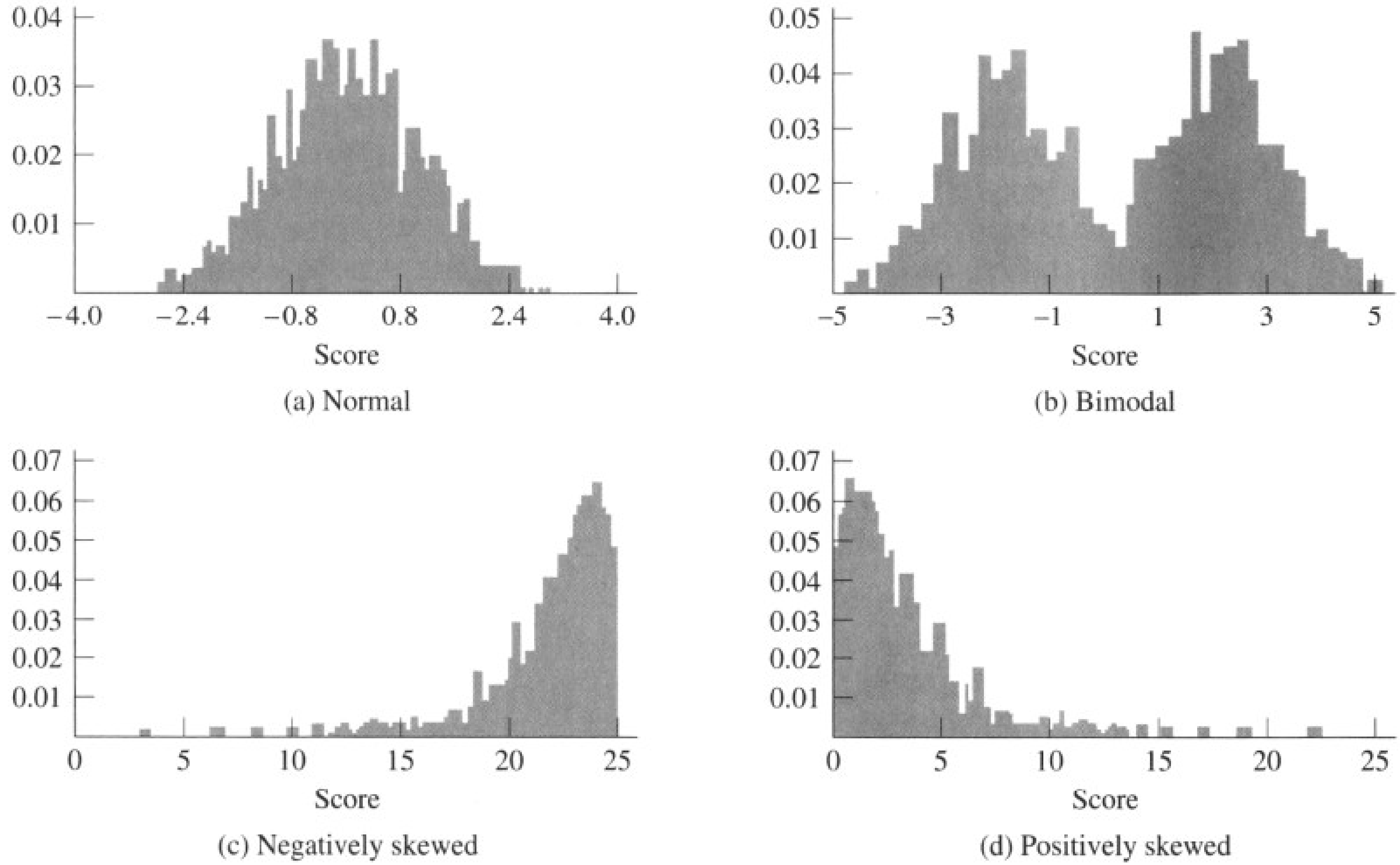
- Descriptive statistics:
  - Describe and explore data
  - All types of graphs, histograms…
    - Understand data distribution
    - Start to think of significance tests
- Inferential statistics:
  - Detect relationships in data
  - Significance tests
  - Infer population characteristics from sample characteristics

Data analysis



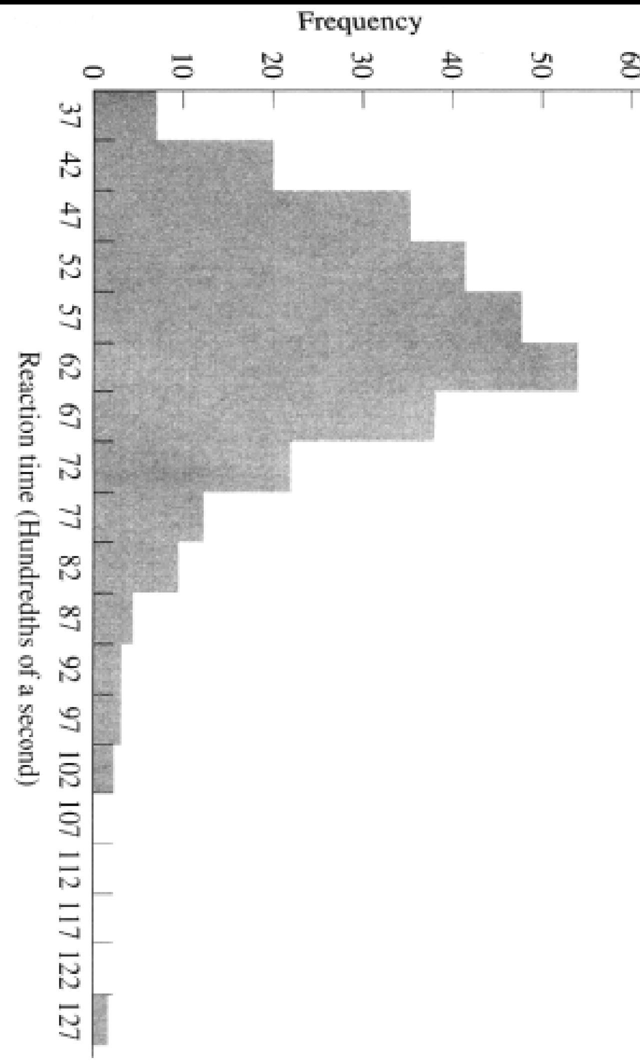
From [Howell 02] p 21

Data analysis



From [Howell 02] p 28

Data analysis



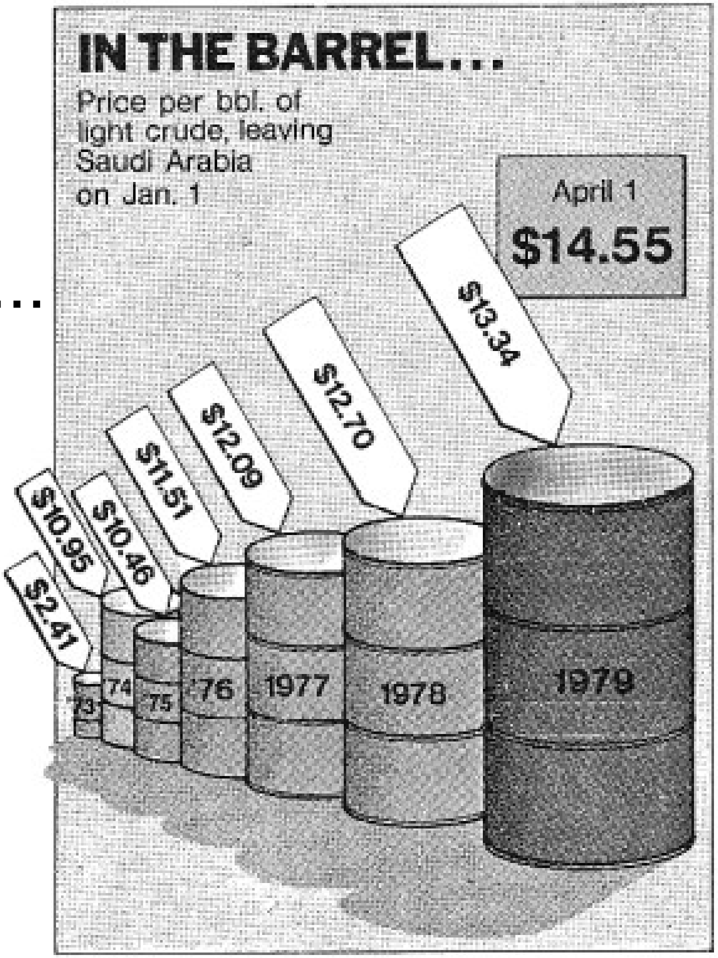
Raw Data	Stem	Leaf
36 37 38 38 39 39 39 40	3s	67
40 40 40 41 41 41 42 42	3.	88999
42 43 43 43 43 43 44 44	4*	0000111
44 44 44 45 45 45 45 45	4t	22233333
45 46 46 46 46 46 46 46	4f	44444555555
46 46 46 46 47 47 47 47	4s	666666666666777777777
47 47 47 47 47 48 48 48	4.	888899999
48 49 49 49 49 49 50 50	5*	00000111111111111
50 50 50 51 51 51 51 51	5t	222222222233333333
51 51 51 51 51 51 51 52	5f	4444445555555
52 52 52 52 52 52 52 52	5s	66666666667777777
52 53 53 53 53 53 53 53	5.	888888888888899999999999
53 54 54 54 54 54 54 55	6*	00000000000011111111111
55 55 55 55 55 55	6t	222222222222222333333333333
	6f	4444444555555555
	6s	666666667777777777777
	6.	889999999
	7*	01111
	7t	2222222333
	7f	44444455
	7s	666677
	7.	88899
	8*	00011
	8t	2333
	8f	5
	8s	67
	8.	8
	9*	0
	9t	
	9f	4455
	9s	
	9.	8
	High	104; 104; 125

FIGURE 2.4 *Stem-and-leaf display for reaction time data*

From [Howell 02] p 21, 23

Data analysis

- Not all examples are good
  - Be fair!!!
  - Tufte has plenty of examples…



MOVING

Data analysis

- Reducing the numbers:
  - Central tendency:
    - Mean, median and mode
  - Variability/dispersion:
    - Mean absolute deviation, variance, standard deviation

Validity of experiments

Mean:

$$\overline{X} = \frac{\sum X}{N}$$

Mean absolute deviation:

$$\text{m.a.d.} = \frac{\sum \left| X - \overline{X} \right|}{N}$$

Variance:

$$s^2 = \frac{\sum \left(X - \overline{X} \right)^2}{N-1}$$

Standard deviation:

$$s = \sqrt{\frac{\sum \left(X - \overline{X} \right)^2}{N-1}}$$

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- **Standard deviation uses same units as samples and mean.**
- **Calculation of population variance σ^2 is theoretical, because μ almost never known and the population size N would be very large (perhaps infinity).**

Validity of experiments

- Hypothesis testing, analysis of variance
  - Read in the PDF, your notes of statistics previous courses…

Pere-Pau Vázquez

IDI – Quantitative and Qualitative Methods for Human-Subject Experiments