Data Science in the Wild

Lecture 6: Running Experiments

Eran Toch

CORNELL TECH

Agenda

1. About experiments
2. Statistical inference
3. Forming hypotheses
4. Designing an experiment

(1) About Experiments

Research Types

- **Observational**: researchers observe what is happening or what has happened in the past and try to draw conclusions
- **Experimental**: researchers impose treatments and controls and then observe characteristic and take measures
    - the researchers manipulate the variables and try to determine how the manipulation influences other variables

Observational Study

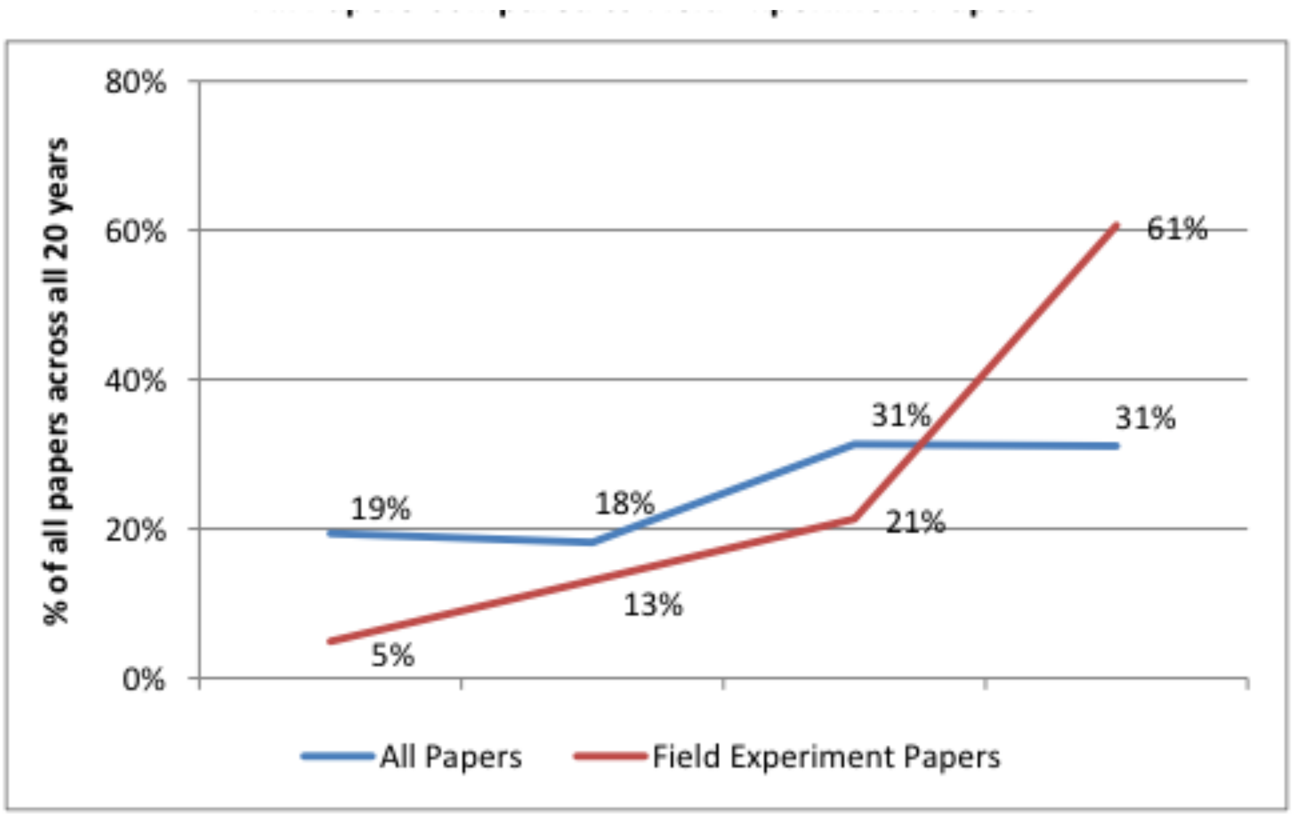
- Are based on observing and recording data
- Associations and predictabilities between variables are analyzed
- Cause and effect are hard (often impossible) to establish
- We cannot test alternatives that do not exist

Experiment Studies

- Are based on a predefined hypothesis
- The experiment design should lead to a clear conformation or rejection of the hypothesis and
- The effect depends solely on conditions which are derived from the hypothesis

Experiments in the Wild

- Experiments are tough
- Some industries were always heavily reliant on experimentation (e.g., pharmaceutical)
- But they are becoming more and more prevalent

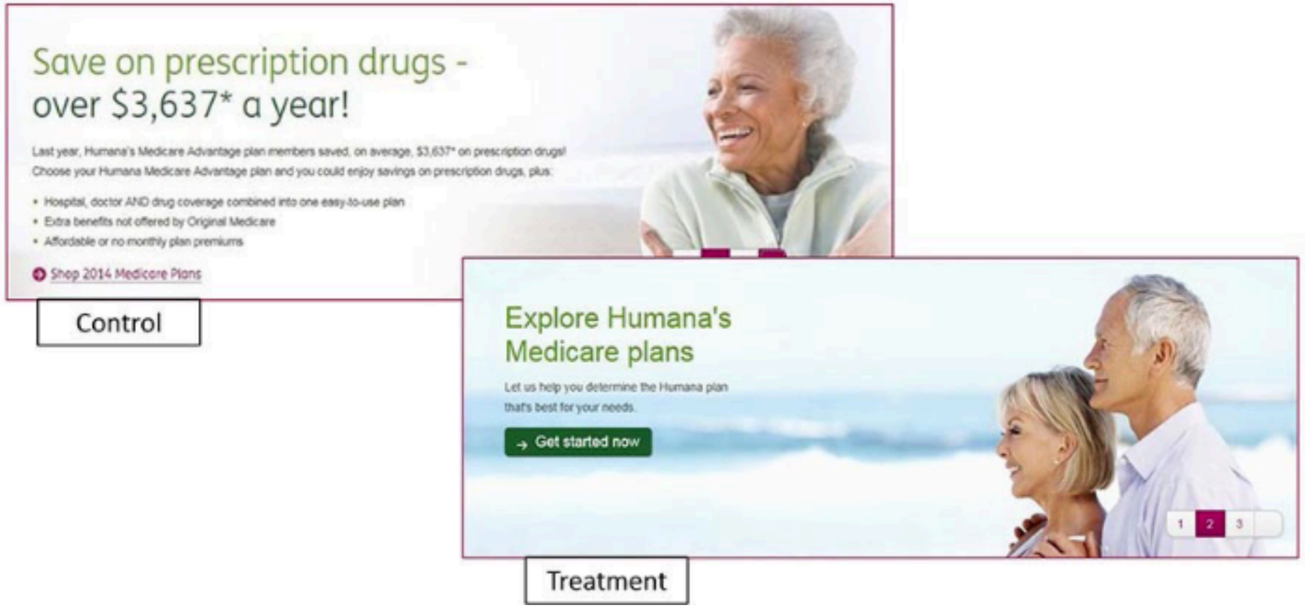


This figure reports a histogram of how many papers were published in each 5-year period. We separately report the findings for all papers, and for just those papers that report findings from a field experiment. The sample size is 3,250 for all papers, and 61 for the field experiment papers. The percentages within each curve add to 100%.

Simester, Duncan. "Field experiments in marketing." *Handbook of Economic Field Experiments*. Vol. 1. North-Holland, 2017. 465-497.

A/B Testing

- A/B testing or split testing is an experimental approach to design
- A portion of the users are presented with the alternative UI
- A better name is multivariate testing (A/B but with more conditions)



Save on prescription drugs - over $3,637* a year!

Last year, Humana's Medicare Advantage plan members saved, on average, $3,637* on prescription drugs! Choose your Humana Medicare Advantage plan and you could enjoy savings on prescription drugs, plus:

- Hospital, doctor AND drug coverage combined into one easy-to-use plan
- Extra benefits not offered by Original Medicare
- Affordable or no monthly plan premiums

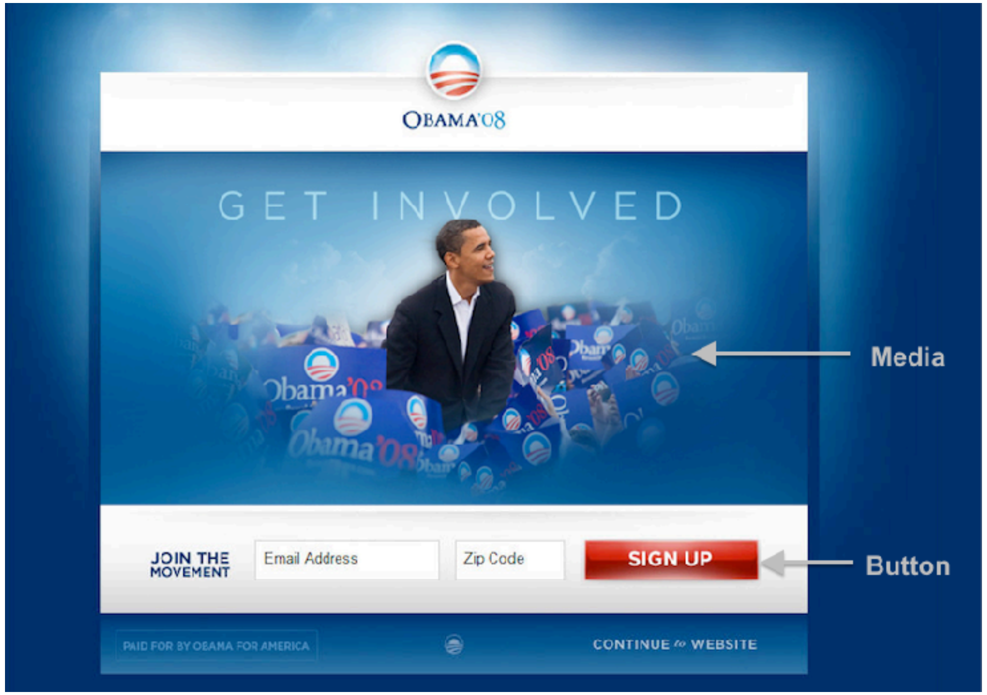
Explore Humana's Medicare plans

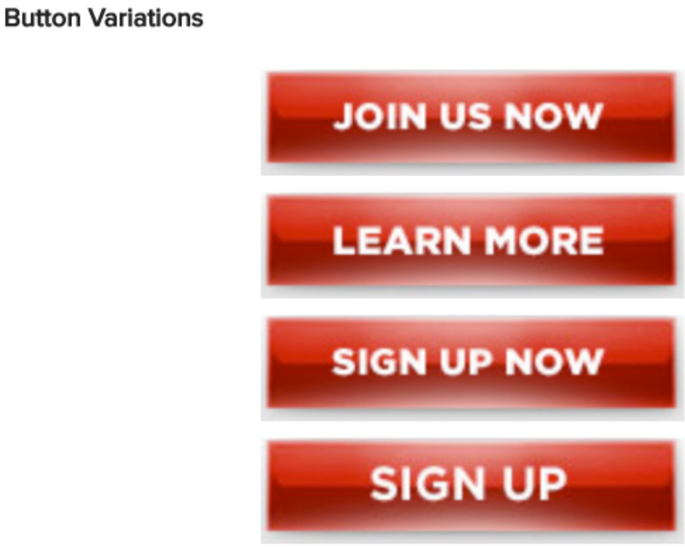
Let us help you determine the Humana plan that's best for your needs.

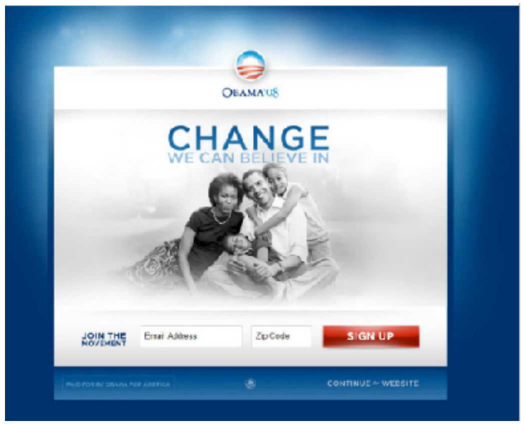
https://www.crazyegg.com/blog/ab-testing-examples/

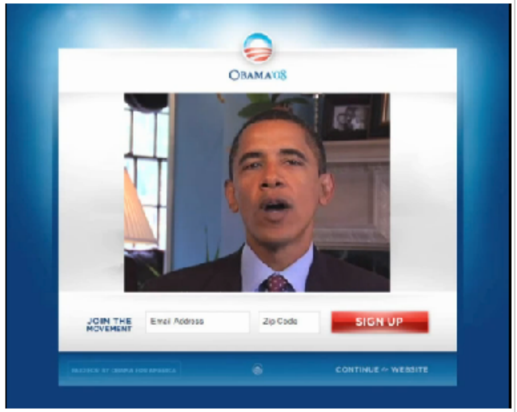
A/B Example

This experiment tested two parts of our splash page: the “Media” section at the top and the call-to-action “Button”

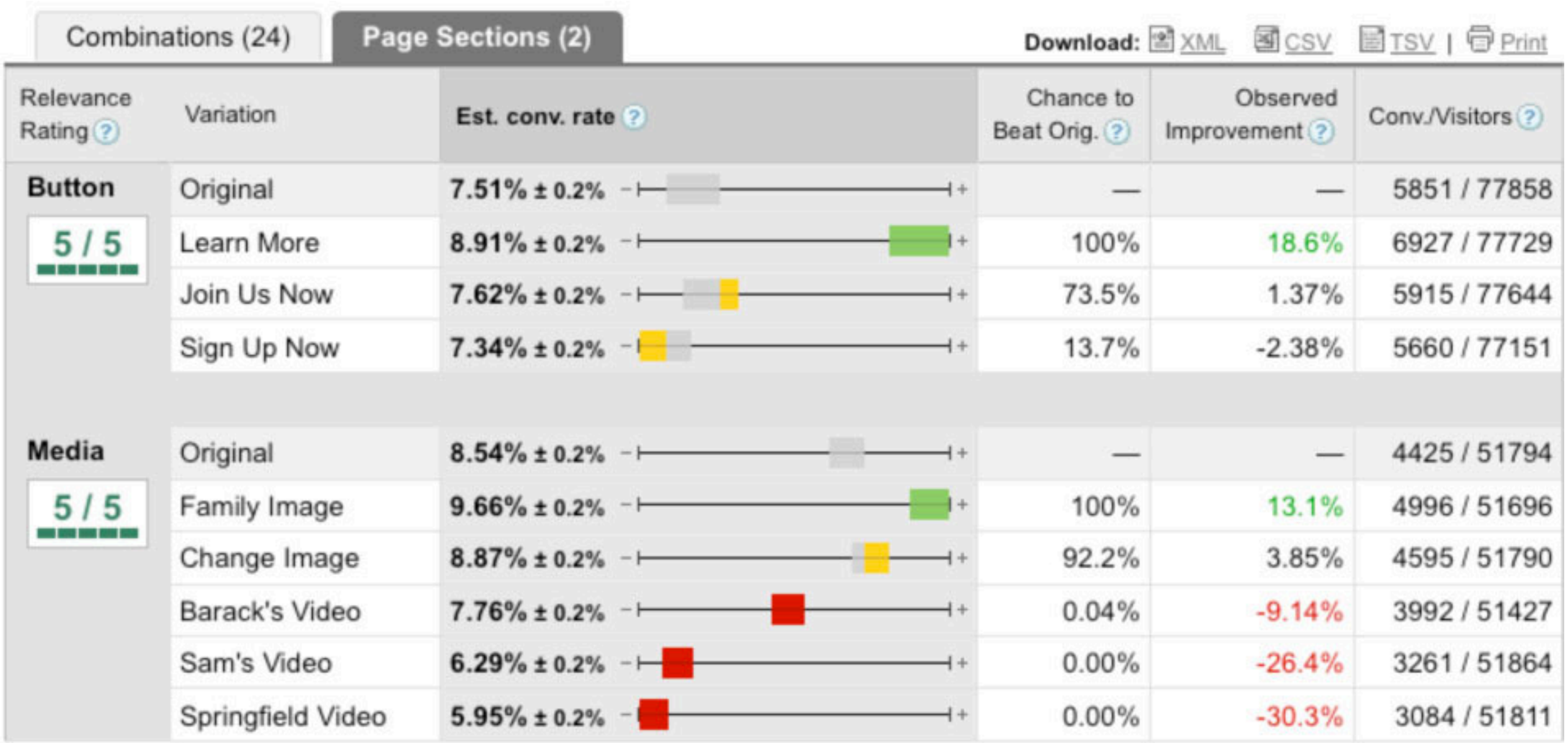




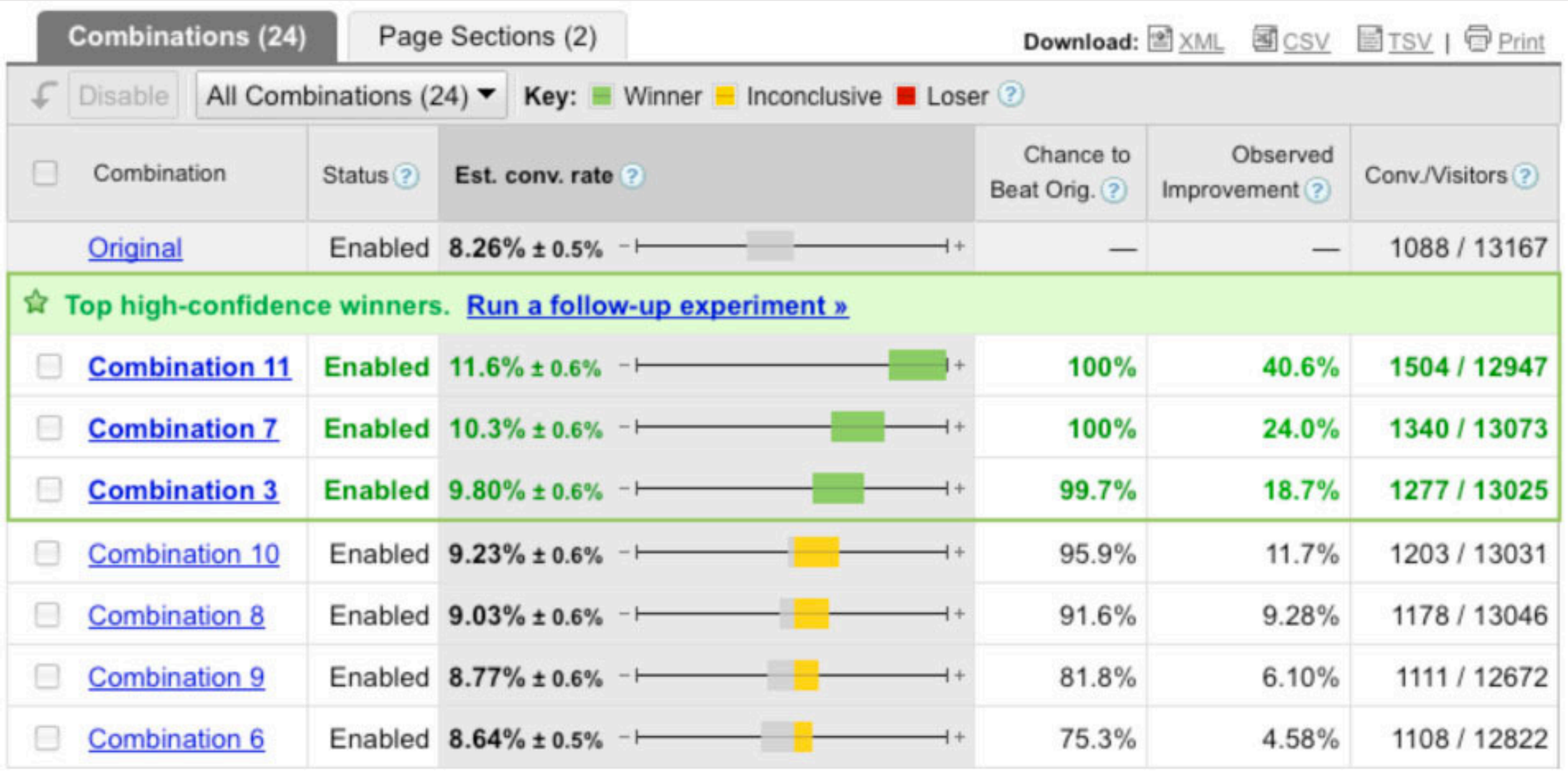




https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/



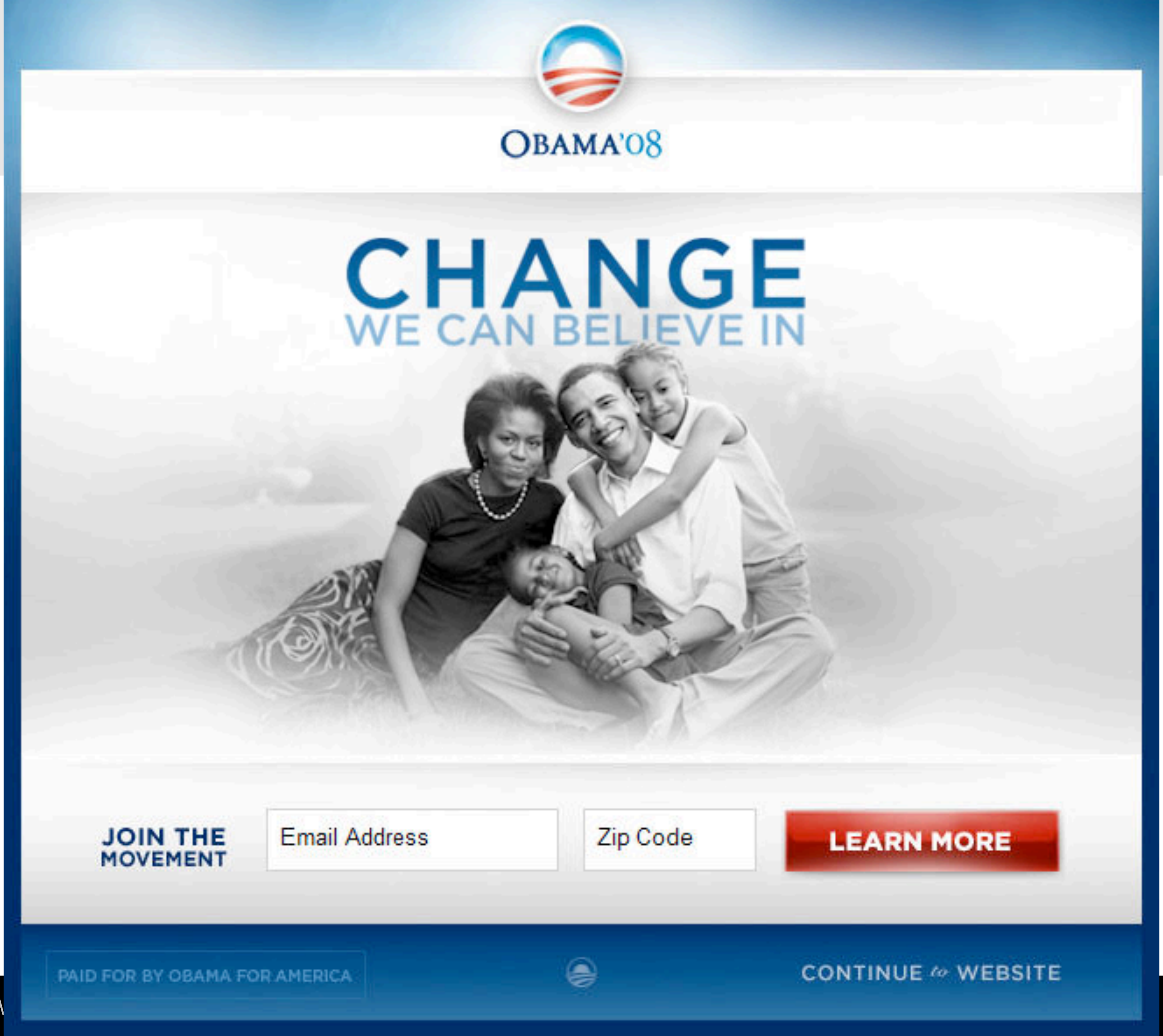
Relevance Rating	Variation	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Button 5 / 5	Original	7.51% ± 0.2%	—	—	5851 / 77858
	Learn More	8.91% ± 0.2%	100%	18.6%	6927 / 77729
	Join Us Now	7.62% ± 0.2%	73.5%	1.37%	5915 / 77644
	Sign Up Now	7.34% ± 0.2%	13.7%	-2.38%	5660 / 77151
Media 5 / 5	Original	8.54% ± 0.2%	—	—	4425 / 51794
	Family Image	9.66% ± 0.2%	100%	13.1%	4996 / 51696
	Change Image	8.87% ± 0.2%	92.2%	3.85%	4595 / 51790
	Barack's Video	7.76% ± 0.2%	0.04%	-9.14%	3992 / 51427
	Sam's Video	6.29% ± 0.2%	0.00%	-26.4%	3261 / 51864
	Springfield Video	5.95% ± 0.2%	0.00%	-30.3%	3084 / 51811



Combinations (24) | Page Sections (2) | Download: XML CSV TSV | Print

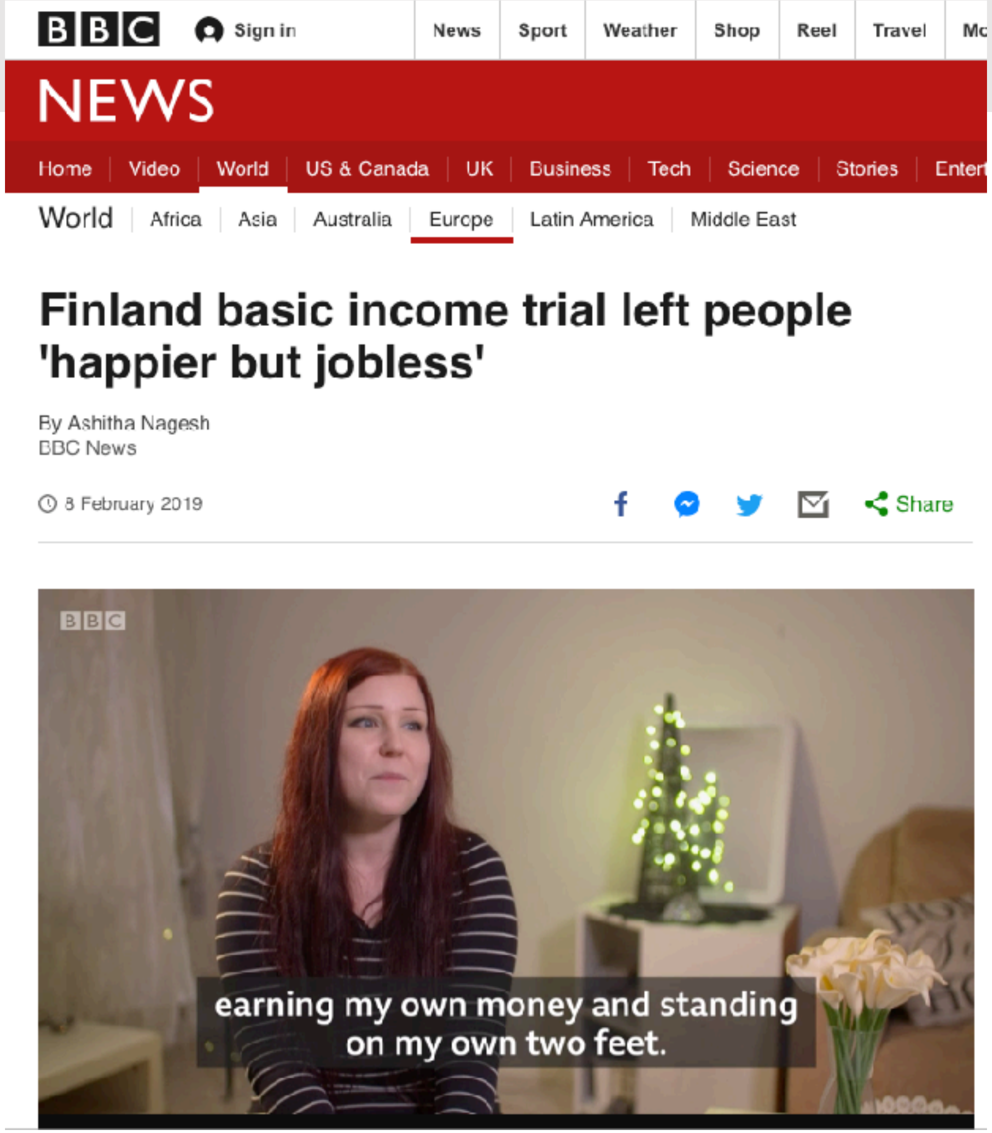
Disable | All Combinations (24) ▾ | Key: 🟩 Winner 🟨 Inconclusive 🟥 Loser ?

□	Combination	Status ?	Est. conv. rate ?	Chance to Beat Orig. ?	Observed Improvement ?	Conv./Visitors ?
	Original	Enabled	8.26% ± 0.5%	—	—	1088 / 13167
☆	Top high-confidence winners. Run a follow-up experiment »					
□	**Combination 11**	**Enabled**	**11.6% ± 0.6%**	**100%**	**40.6%**	**1504 / 12947**
□	**Combination 7**	**Enabled**	**10.3% ± 0.6%**	**100%**	**24.0%**	**1340 / 13073**
□	**Combination 3**	**Enabled**	**9.80% ± 0.6%**	**99.7%**	**18.7%**	**1277 / 13025**
□	Combination 10	Enabled	9.23% ± 0.6%	95.9%	11.7%	1203 / 13031
□	Combination 8	Enabled	9.03% ± 0.6%	91.6%	9.28%	1178 / 13046
□	Combination 9	Enabled	8.77% ± 0.6%	81.8%	6.10%	1111 / 12672
□	Combination 6	Enabled	8.64% ± 0.5%	75.3%	4.58%	1108 / 12822


OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE

Experiments in the Wild

- Finland has begun reporting on its two-year experiment with a guaranteed monthly cash for citizens.
- The program involved a couple of thousand unemployed Finns between the ages of 25 and 58, who got €560 ($634) a month through 2017 and 2018 instead of basic unemployment benefits.
- The results were compared with a control group with the same characteristics



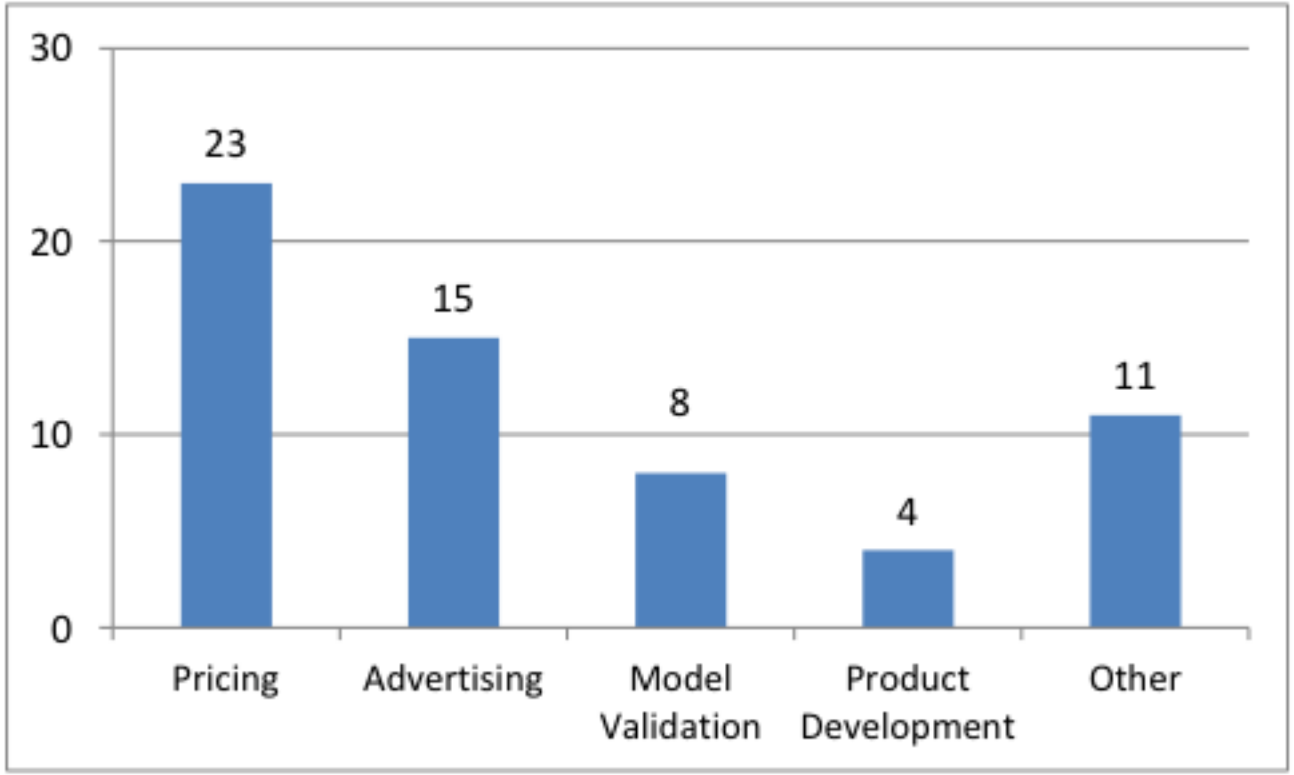
Finland basic income trial left people 'happier but jobless'

By Ashitha Nagesh
BBC News

8 February 2019

earning my own money and standing on my own two feet.

What do companies experiment with?

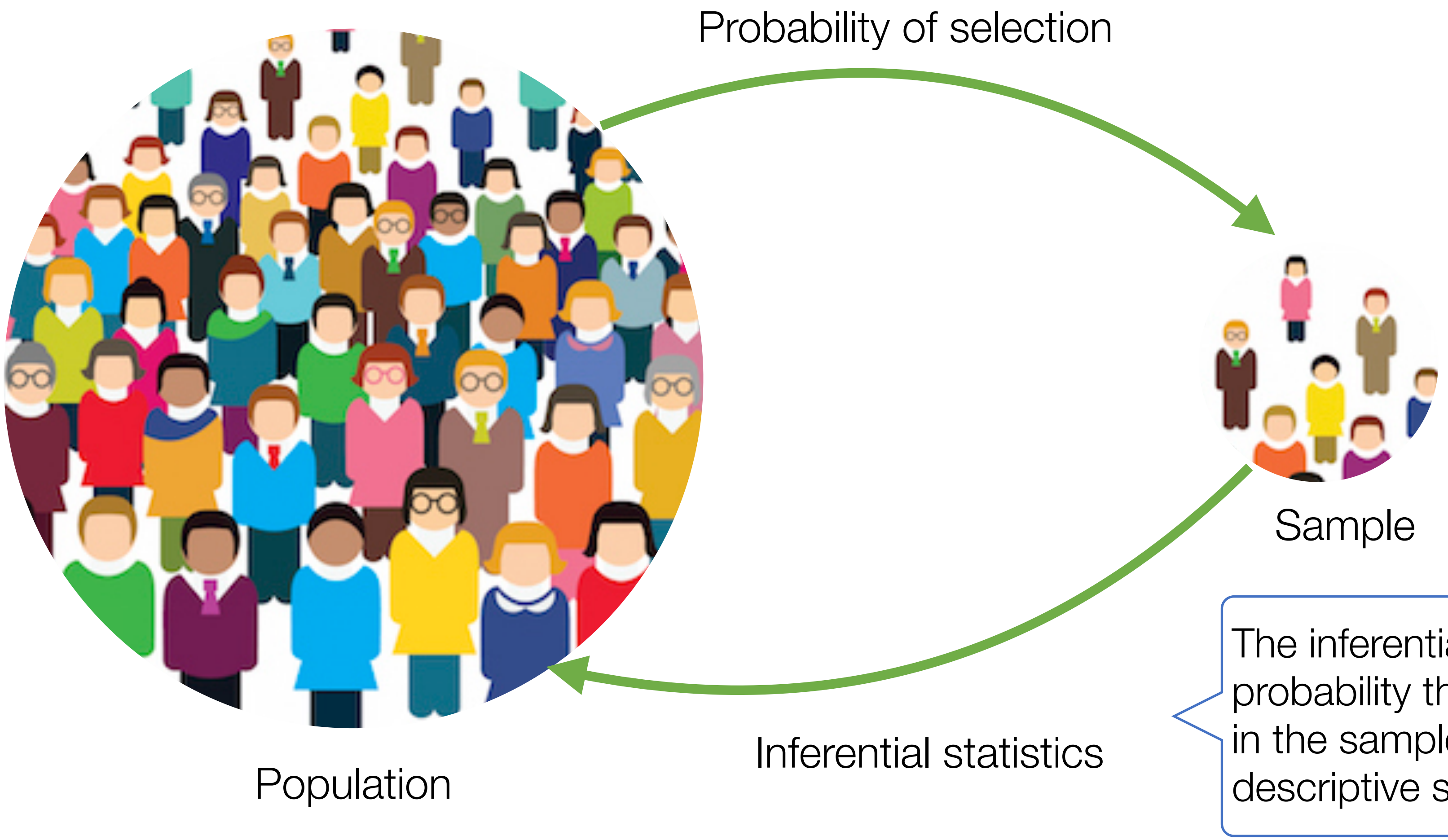


This figure reports the number of published papers by topic. The sample size is 61 papers.

Simester, Duncan. "Field experiments in marketing." *Handbook of Economic Field Experiments*. Vol. 1. North-Holland, 2017. 465-497.

(2) Statistical Inference

Statistical Inference



Probability of selection

Sample

Population

Inferential statistics

The inferential statistics reflect the probability that the descriptive statistics in the sample will be correlated with the descriptive statistics in the population

Observation vs. Experimentation

Example:

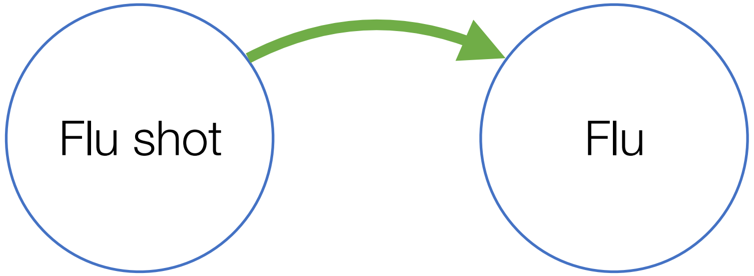
20 people went for a flu shot to a public hospital

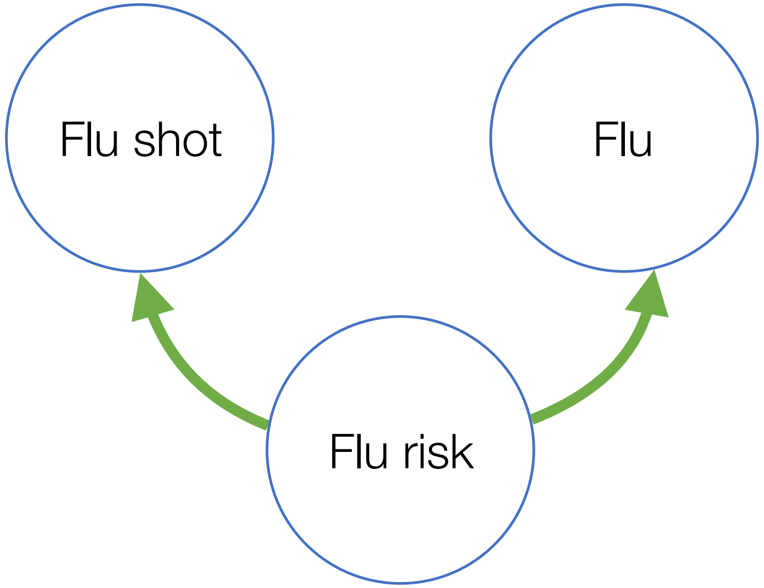
After a month, an independent researcher checked how many of them got flu

7 of them got flu, and the others didn’t

The Problem with Causation

- Which conclusions can we derive from case 1?
  - Flu shots increase the probability of flu?
  - Flu shots decrease the probability of flu?
- **Confounding** factors

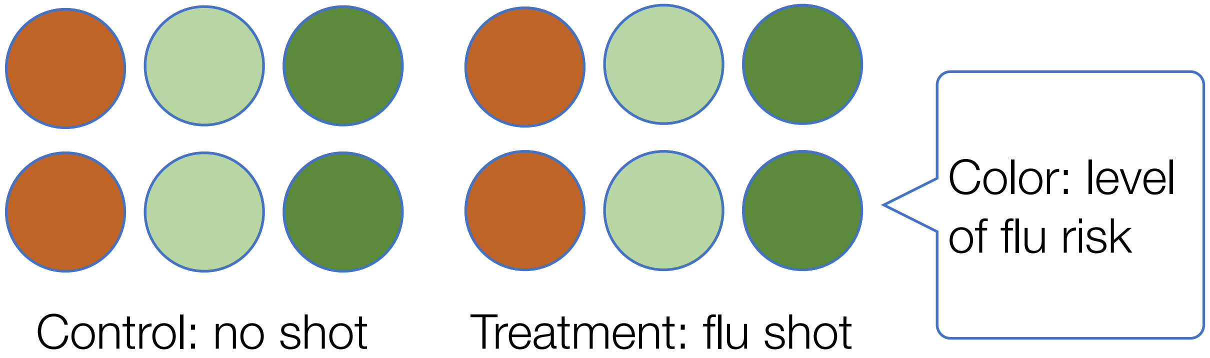




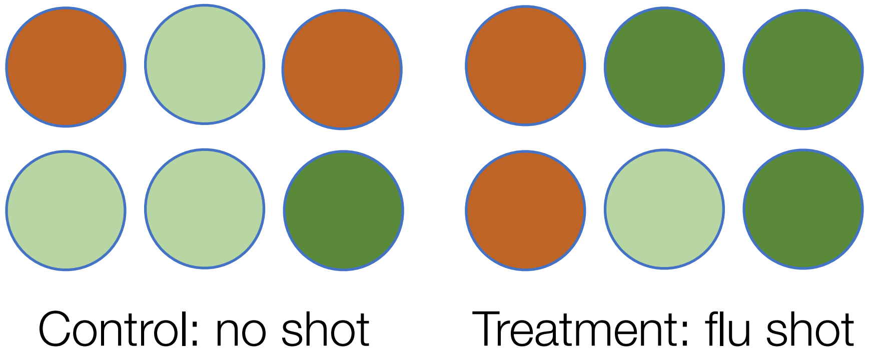
Dealing with cofounding factors

Experimentation enables the identification of casual relations (X is responsible for Y) by **trying** to control all interfering variables

Stratify the variables: make sure every condition has the same values of stratifying variables



Randomize the variables: randomly assign participants (data points) to conditions

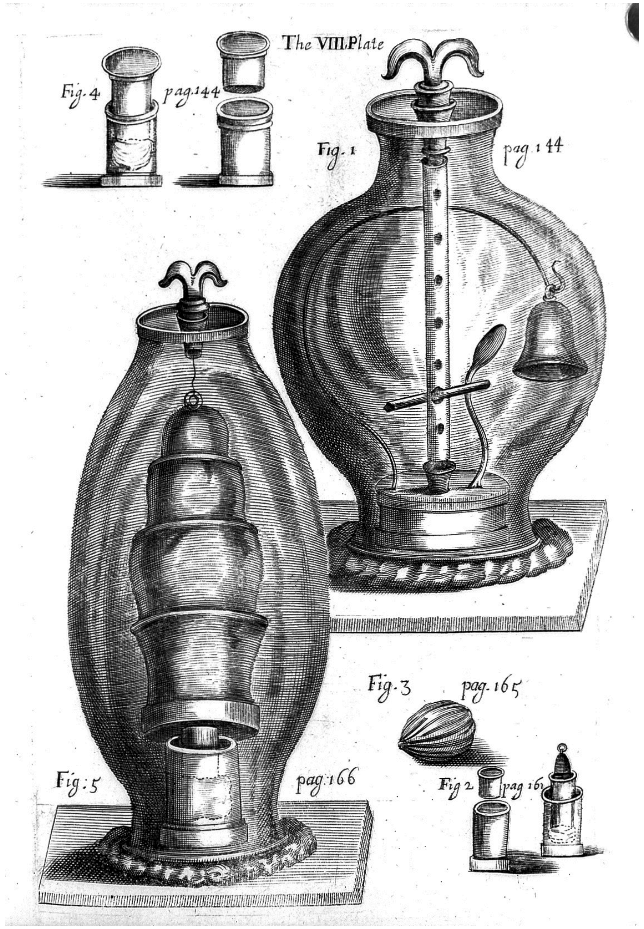


Finding Causation

- **Example 2**: We randomly select 20 people with similar health condition, and randomly assign them to two groups: A, and B
- Then, we give the flu shots to group A, and placebo to group B, and observe how many got flu after a month

Issues with experiments

- Forming hypotheses
- Experimental design
    - Power analysis
- Experimental analysis
    - Parametric tests
    - Non-parametric tests
- Reproducibility



(4) Designing Experiments

Hypothesis

- An experiment normally starts with a research hypothesis
- A hypothesis is a precise problem statement that can be directly tested through an empirical investigation
- In most cases, a hypothesis describes the effect of some **treatment**
- Compared with a theory, a hypothesis is a smaller, more focused statement that can be examined by a single experiment

Where do Hypotheses Come From?

- Business question
- A phenomenon which is unexplained by a theory
- A phenomenon which contradicts an established theory
  - ★ I.e., Rationality in economic decision making
- Contradictions within a theory

Types of Hypotheses

1. **Null hypothesis - H_0**
   - States the numerical assumption to be tested
   - Reflects no effect of the treatment

2. **Alternative hypothesis - H_A**
   - The opposite of the null hypothesis
   - Reflects some effect of the treatment
   - Generally, the goal of an experiment is to find statistical evidence to refute or nullify the null hypothesis in order to support the alternative hypothesis

One / two tailed hypotheses

- Given some statistics about two samples (let's say mean), μ_1 and μ_2
- Two tailed hypothesis is not directional, and they mean that the two statistics are taken from the same population:

  H_0: $\mu_1 = \mu_2$

- A one-tailed hypothesis (tested using a one-sided test) is an inexact hypothesis in which the value of a parameter is specified as being either:

  H_0: $\mu_1 - \mu_2 \leq 0$

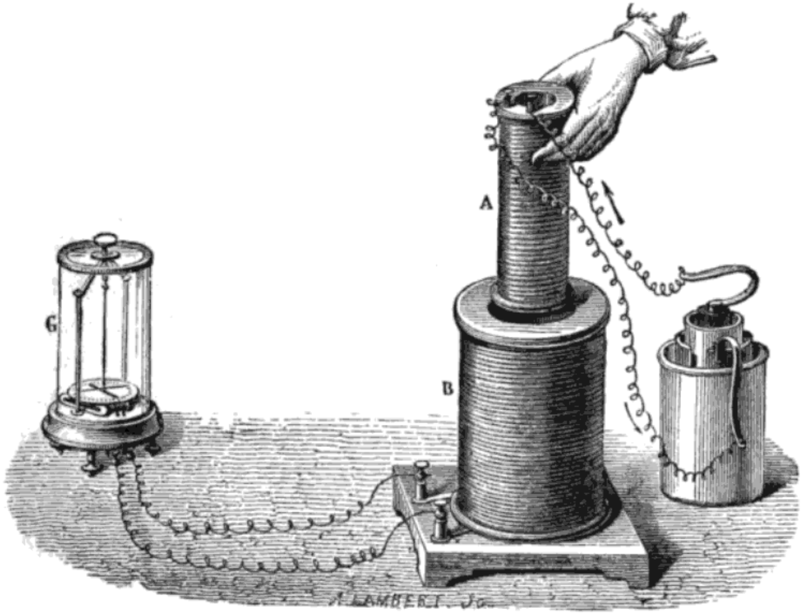
  H_A: $\mu_1 - \mu_2 > 0$

Experimental Design

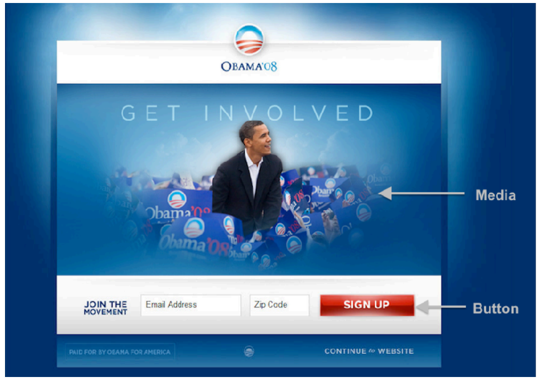
- Experimental design should help us accept either of the hypotheses
- It should show internal validity
    - That we measure our actual hypothesis
- And also the external validity
    - That what we’ve learned is also true for the actual world

Components of Experiments

- **Units**: the objects to which we apply the experiment treatments. In human-based research, the units are normally human subjects with specific characteristics, such as gender, age, or computing experience
- **Conditions**: the different treatments that we test
- **Assignment method**: the way in which the experimental units are assigned different treatments
- **Variables**: the elements that we measure

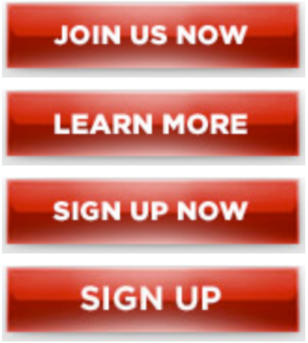


Example



- Units: 2000 site visitors
- Conditions: 4 types of buttons
- Assignment method: random assignment of site visitors to the experiment and then random assignment to the 4 conditions with uniform distribution
- Measures: measuring age, state, conversion rate and time on the site

Button Variations



Variables

- **Independent variables (IV)** refer to the factors that the researchers are interested in studying or the possible “cause” of the change in the dependent variable
  - IV is independent of what will happen in the experiments
  - Conditions are generally seen as IV

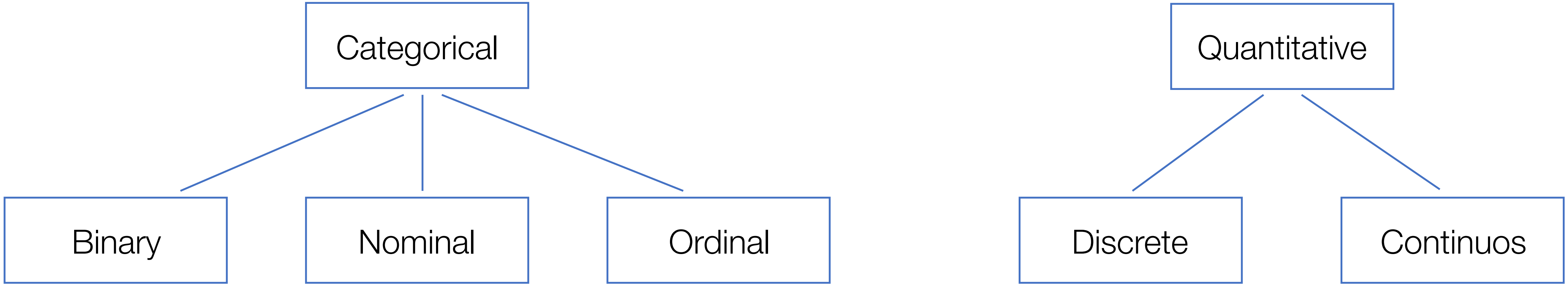
- **Control variables** are independent variables that are kept constant throughout the experiment

- **Dependent variables (DV)** refer to the outcome or effect that the researchers are interested in
  - DV is dependent on a participant’s behavior or the changes in the IVs
  - DV is usually the outcomes that the researchers need to measure

Typical Dependent Variables

- Conversion rate
- Revenue
- Survival
- Drug efficiency
- Accuracy (e.g., error rate)
- Subjective satisfaction
- Ease of learning and retention rate
- Physical or cognitive demand (e.g., NASA task load index)
- Social impact of the technology.

Types of data



Categorical

Binary

2 categories

Nominal

Many categories

Ordinal

Many categories
and order matters

Quantitative

Discrete

Numerical

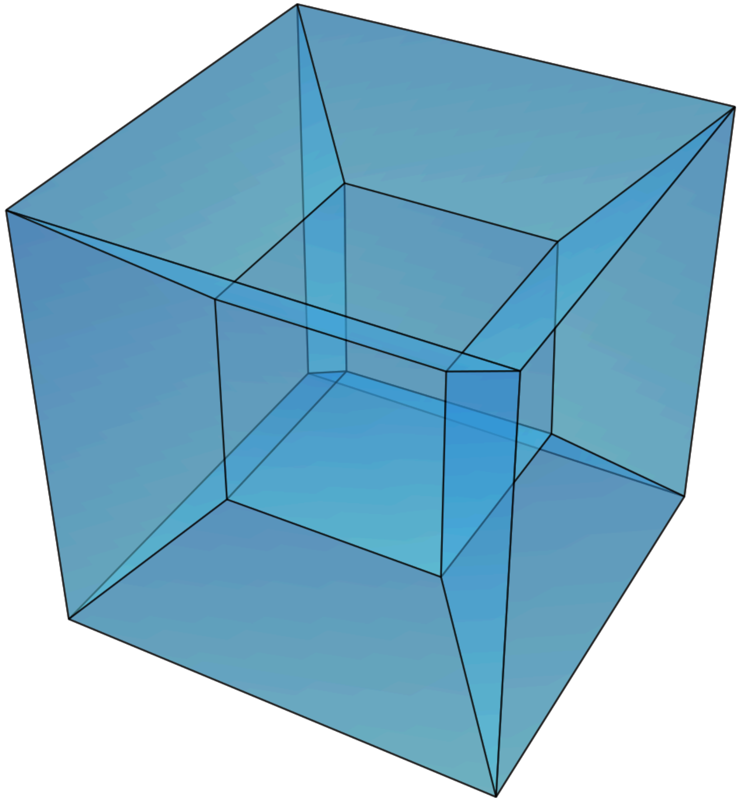
Continuos

Uninterrupted

http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

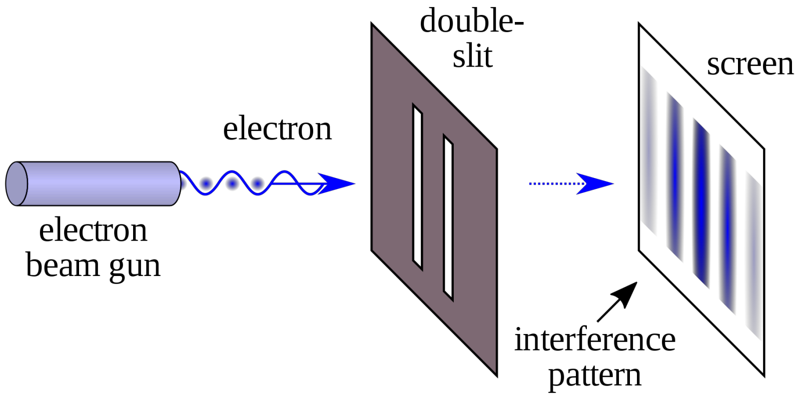
Data Dimensionality

- **Univariate**: Measurement made on one variable per participant
- **Bivariate**: Measurement made on two variables per participant
- **Multivariate**: Measurement made on many variables per participant



Basic design questions

- How many independent variables do we want to investigate in the experiment?
- How many different values does each independent variable have?
- Can we identify effects?
- Interaction between variables



Basic Design Structure

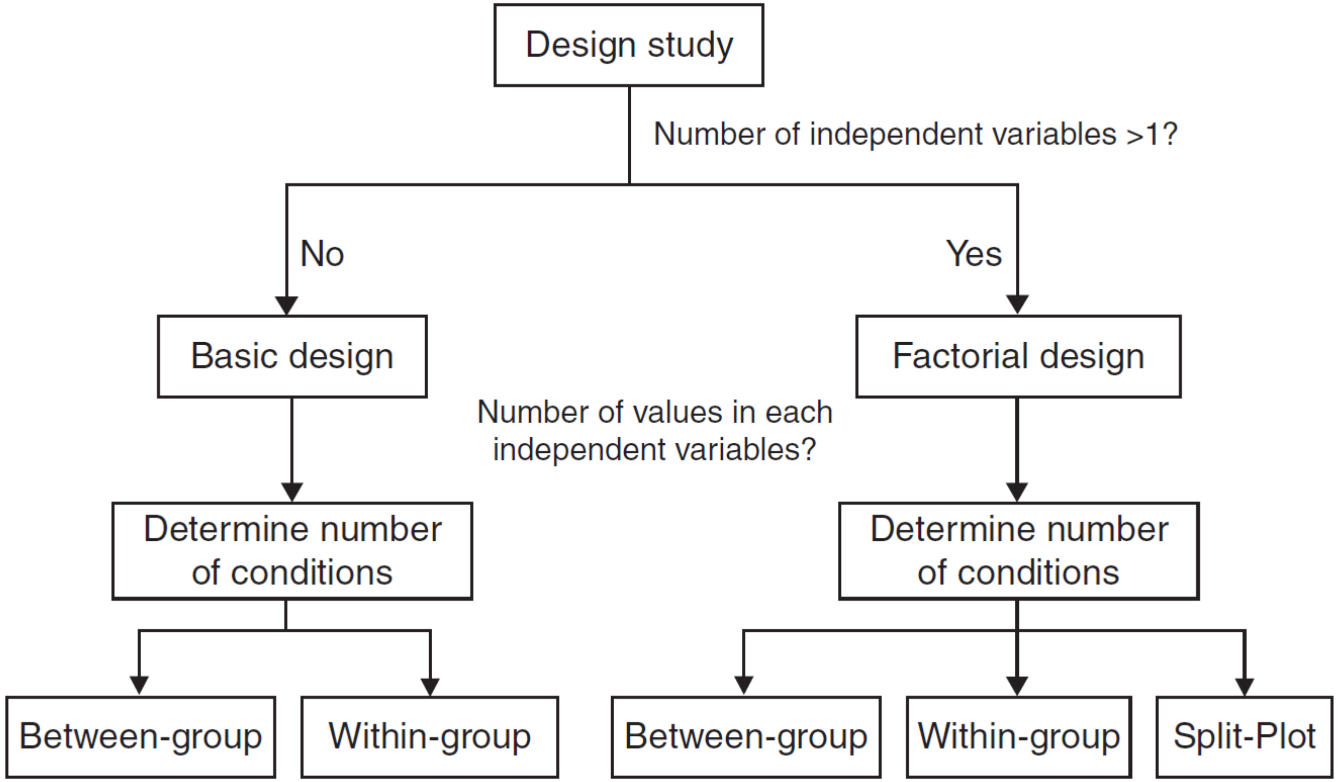
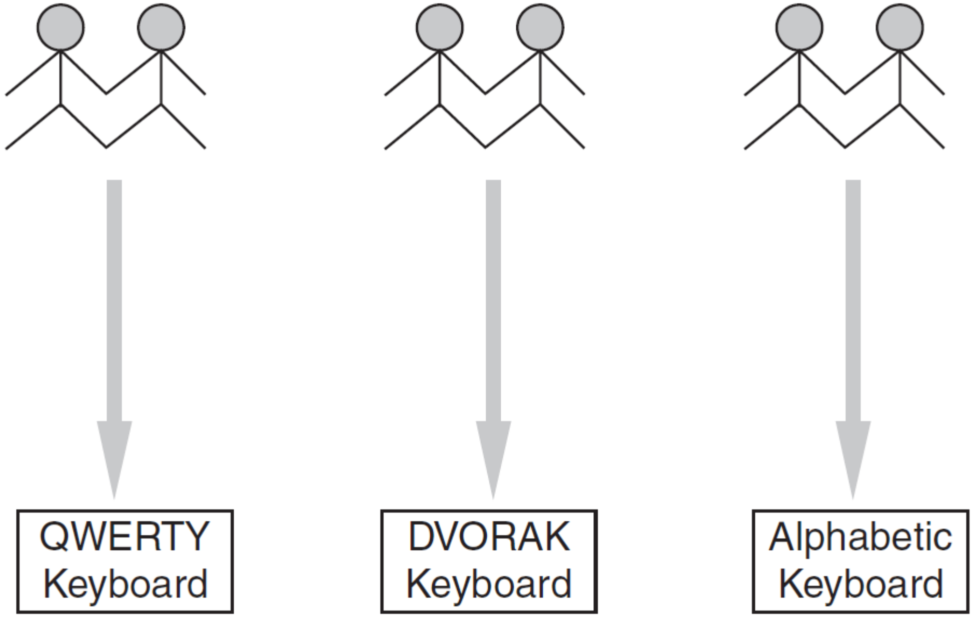

Design study
Number of independent variables >1?
No
Yes
Basic design
Factorial design
Number of values in each independent variables?
Determine number of conditions
Determine number of conditions
Between-group
Within-group
Between-group
Within-group
Split-Plot

Figure 3.2 Determining the experiment structure.

Single independent variable

Between Group Design (single value)

- Investigating one independent variable
- One participant only experience one condition
- Also called “between subject design”

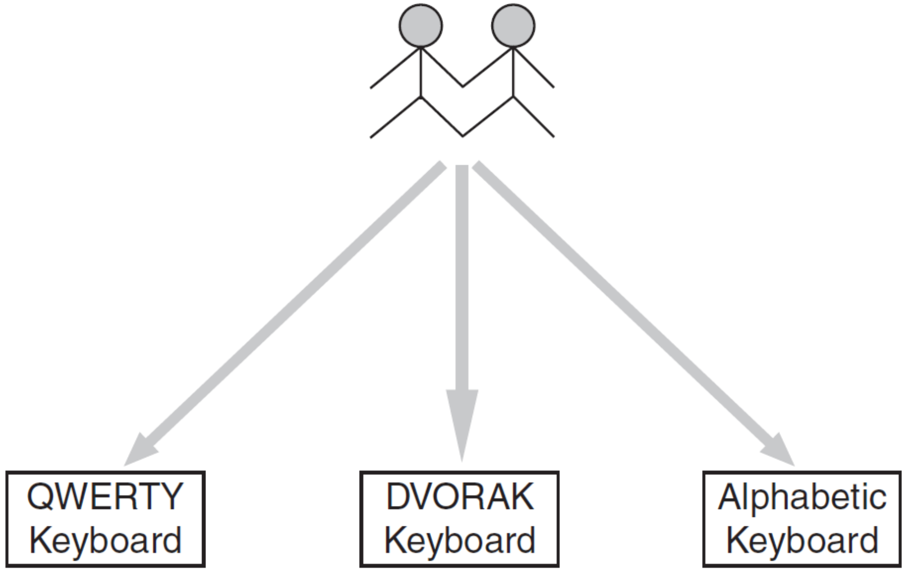


Between Group Design

- **Advantages**
  - Cleaner, better control of learning effect
  - Requires shorter time for participants
  - less impact of fatigue and frustration
- **Disadvantages**
  - Impact of individuals difference
  - Harder to detect difference between conditions
  - Require larger sample size

Within Group Design (single value)

- Investigating one independent variable
- Also called ‘within subject design’
    - or repeated-measures design
- One participant experience multiple conditions



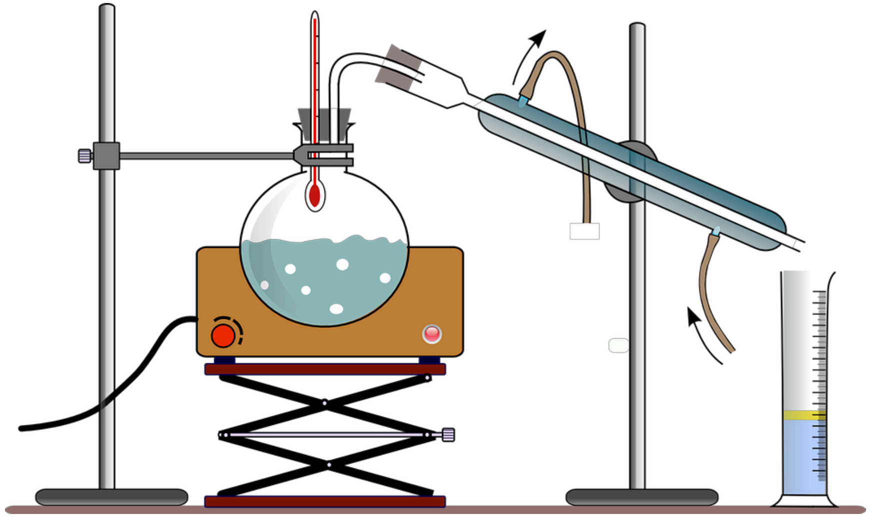
Within-group design

- Advantages
  - Requires smaller sample size
  - Easier to detect difference between conditions
  - Variance due to participants predispositions will be approximately the same across test conditions
  - No need to balance groups of participants

- Disadvantages
  - Order effects
  - Takes longer time
  - Larger impact of fatigue and frustration

Order effects

- **Learning**: The participants learn a skill at the first condition and carry it out to the next
- **Fatigue**: participants' performance will become worse on conditions that follow other conditions
- **Interference**: More generally, any type of interference that is related to the order of the conditions



Combating order effects

- How to fight confounding factors?
    - Randomization the order of experimental conditions
    - Providing training time to avoid the learning curve
    - Reducing the time it takes to complete an assignment

Latin Square Design

- A Latin square is an n*n table filled with n different symbols in such a way that each symbol occurs exactly once in each row and exactly once in each column
- **Counterbalancing**: the order of presentation is different for each group of participants, the learning effect tends to balance out

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

unbalanced: 3 before is over sampled

$$\begin{bmatrix} a & b & d & c \\ b & c & a & d \\ c & d & b & a \\ d & a & c & b \end{bmatrix}$$

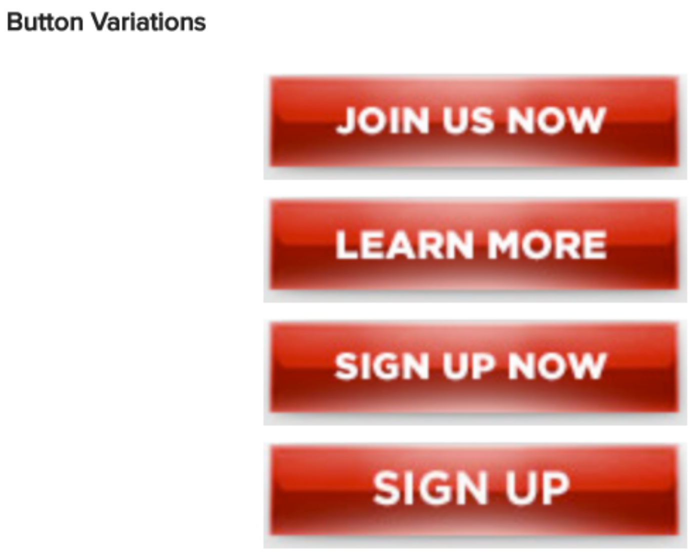
Balanced: no combination has higher probability

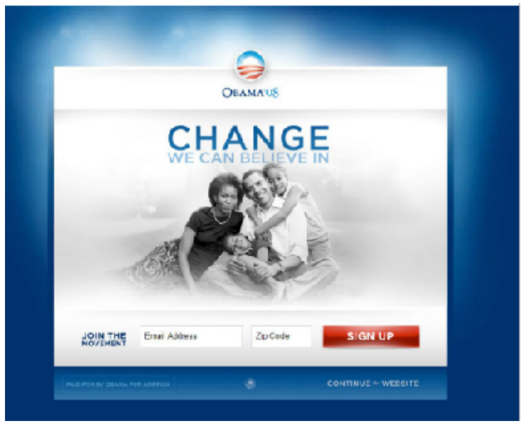
Summary: Between group vs. Within group

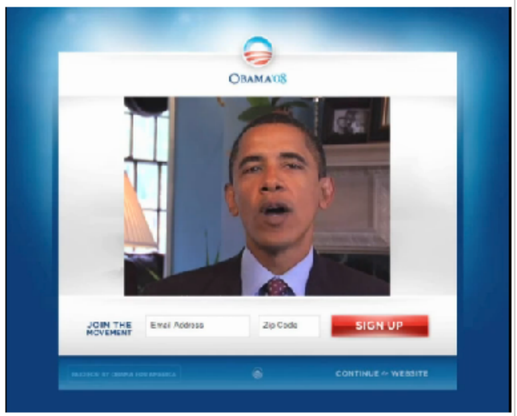
- **Between-group design should be taken when:**
  - Simple tasks
  - Learning effect has large impact
  - Within-group design is impossible
- **Within-group design should be taken when:**
  - Learning effect has small impact
  - Small participant pool

Factorial Design

What happens when we have more than one variable?







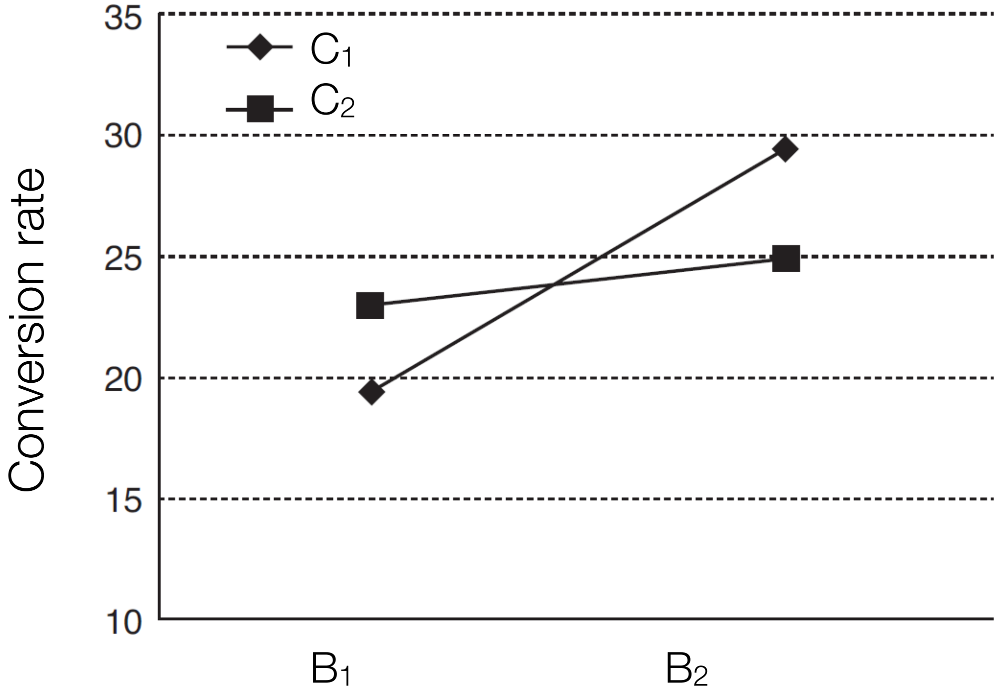
More than one independent variable

- Factorial design divides the experiment groups or conditions into multiple subsets according to the independent variables
    - Each independent variable value is called a factor
    - Each value is studied in interaction with other values
- Thus, we can study interaction effects as well as the impact of independent variables

	B_1 - Button 1	**B_1 - Button 2**
C_1 - Content 1	C_1, B_1	C_1, B_2
C_1 - Content 2	C_2, B_1	C_2, B_2

Interaction effect

Many times, we want to study the interaction effect: the effect of one independent variable on the dependent variable, depending on the particular level of another independent variable



Number of Conditions

- Number of conditions, where C is the number of conditions, V is the number of levels in each variables.

$$C = \prod_{a=1}^{n} Va$$

- Imagine we want to compare three types of content and the effect of two types of buttons
  C = 3 * 2 = 6

Design Options

- Three options of factorial design
  - Between group design
  - Within group design
  - Split-plot design

- Split-plot design
  - Has both a between-group and a within-group component
  - Is multi-factored (each factor is a variable)

Limitations of Experimental Research

- Experimental research requires well-defined, testable hypotheses that consist of a limited number of dependent and independent variables
- Experimental research requires strict control of factors that may influence the dependent variables
- Experiments may not be a good representation of users' typical interaction behavior:
    - External validity is a key
    - As well as sampling

Summary