Econometrics of Experiments

Wang, Yu

dr.yuwang@outlook.com

December 24, 2024

Overview

1. Key Concepts and Definitions
2. Treatment Testing
3. Power Analysis
4. Clustering vs. Multilevel Modelling
5. Data Types
6. Meta-Regression

Table of Contents

Key Concepts and Definitions

The experimenter is interested in the value of a parameter θ.

- The experiment has two treatments, treatment 1 (“control”) and treatment 2.
- This parameter has (population) value θ_1 under treatment 1 and θ_2 under treatment 2.

The “true effect size” δ is the difference between the parameter values under the two treatments, i.e., $\delta = \theta_2 - \theta_1$.

Key Concepts and Definitions

A treatment test has a null hypothesis H_0 and an alternative hypothesis H_1.

- H_0: $\delta = 0$ (the true effect size is zero).
- 1) H_1: $\delta > 0$ (one-sided alternative and one-tailed test).
- 2) H_1: $\delta < 0$ (one-sided alternative and one-tailed test).
- 3) H_1: $\delta \neq 0$ (two-sided alternative and two-tailed test).

One-sided alternatives are proposed when the researcher has a **prior belief** about the direction of the effect, coming from economic theory.

Key Concepts and Definitions

The hypothesis testing:

- 1) Compute the test statistic which is a function of the n data values in the sample (n is the sample size).
- 2) Compare the test statistic to the null distribution (i.e., the distribution that the statistic would in theory follow if the true effect size were zero).
- 3) The tails of this distribution form the rejection region of the test:

  The rejection region is determined by whether the test is two-tailed or one-tailed, and by the chosen size α of the test.

  The point at which the rejection region starts is the critical value of the test.

  If the test statistic falls in this region, the null hypothesis is rejected; if the test statistic falls elsewhere, the null hypothesis is not rejected.

Key Concepts and Definitions

The p-value of the test is the probability of obtaining a test statistic that is at least as extreme as the one obtained under the assumption that H_0 is correct.

- It allows a conclusion to be drawn without comparing a test statistic to a critical value.
- It is a measure of the strength of evidence against the null (i.e., the strength of evidence of an effect).
- If $p < 0.10$, there is mild evidence of an effect; if $p < 0.05$, evidence; if $p < 0.01$, strong evidence; if $p < 0.001$, overwhelming evidence.

Prior beliefs are very useful because they can be used to boost the chances of obtaining a conclusive result.

- For a one-tailed test based on a prior belief, the p-value is half of the p-value for the corresponding two-tailed test.
- Therefore one-tailed tests are more likely to find evidence of an effect.

Key Concepts and Definitions

Two types of errors:

- Type 1 error (false positive): reject H_0 when it is true.
- **The size of the test** α: the probability of a type 1 error.
- Type 2 error (false negative): fail to reject H_0 when it is false.
- β: the probability of a type 2 error.

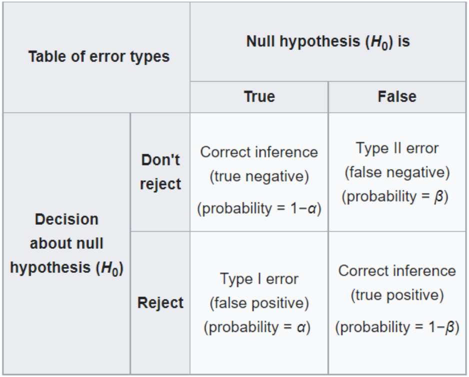
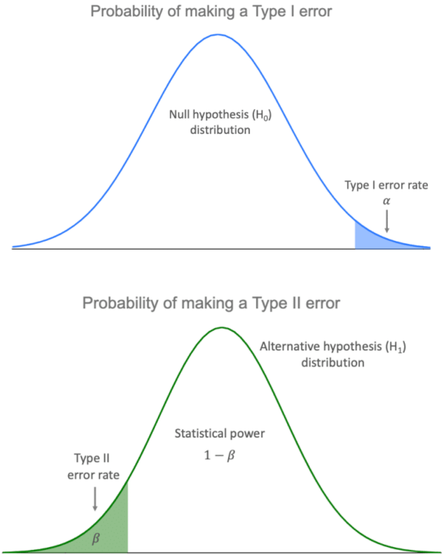


Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Key Concepts and Definitions



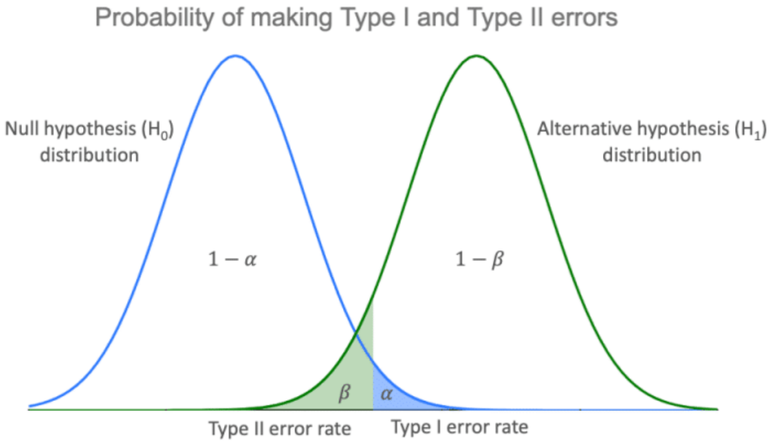
Probability of making a Type I error

Probability of making a Type II error

Null hypothesis (H_0) distribution
Type I error rate
α
Alternative hypothesis (H_1) distribution
Statistical power
$1-\beta$
Type II error rate
β

Key Concepts and Definitions

Trade-off between type 1 errors and type 2 errors:



Key Concepts and Definitions

The power of a test $\pi = 1 - \beta$ is the probability of rejecting the null hypothesis when it is false.

- The power of a test is determined by

  1) the true effect size δ,

  2) the sample size *n*,

  3) whether the test is one-tailed or two-tailed,

  4) the chosen value of size α, and

  5) the chosen statistical test.

- The higher the chosen size of the test α is, the higher the probability of type 1 error, which has the benefit of higher power π.

Power analysis is used to compute the power π of a given test, or to find the sample size *n* required to meet a given power requirement.

Key Concepts and Definitions

The choice of size α depends on the type of hypothesis (or research question):

- 1) Case 1: a crime suspect is guilty (or innocent).
- 2) Case 2: a patient is suffering from a contagious disease (or healthy).

Type 1 errors or type 2 errors, which are more serious? A high level of α or a low level α, which will you choose?

Key Concepts and Definitions

Economic conventions:

- Size $\alpha = 0.05$;
- Power $\pi = 0.80$ and $\beta = 1 - \pi = 0.20$ (no formal standards for adequate power).

If so, then Pr(type 2 errors) : Pr(type 1 errors) $\approx 4 : 1$.

Key Concepts and Definitions

Consider the cases:

- 1) Ideally, what is the proportion of false positive publications in an economic journal?
- 2) If N independent experimenters work on a same topic of which the true effect size is zero, what is the probability that at least one of them find a statistically significant result (and submit to the journal)?
- 3) ... the true effect size is non-zero, what is the probability that none of them find a statistically significant result?

Robustness is important to economic journals and professions.

Key Concepts and Definitions

Replication crisis: many scientific studies are difficult or impossible to replicate or reproduce.

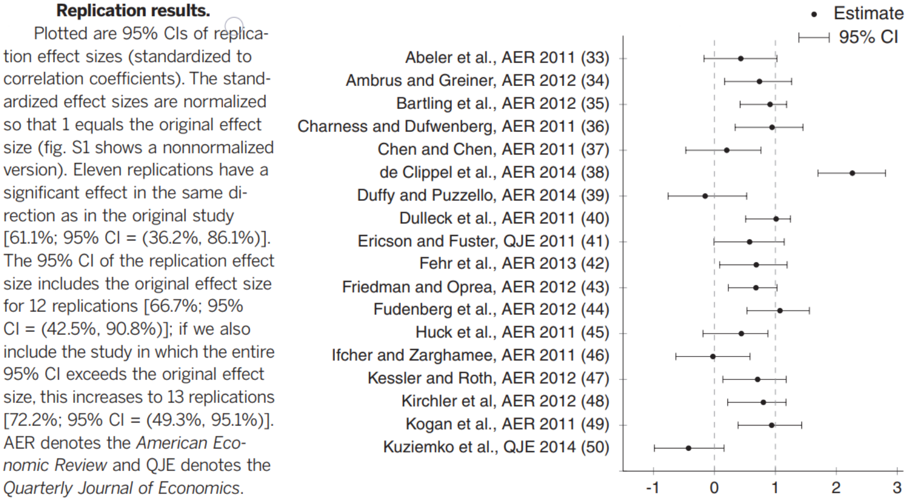
- Camerer et al. (2016)

  One-third of 18 experimental studies from two top-tier economics journals (*AER* and *QJE*) failed to replicate.

- Ioannidis et al. (2017)

  "The majority of the average effects in the empirical economics literature are exaggerated by a factor of at least 2 and at least one-third are exaggerated by a factor of 4 or more."

Key Concepts and Definitions



Replication results. Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized so that 1 equals the original effect size (fig. S1 shows a nonnormalized version). Eleven replications have a significant effect in the same direction as in the original study [61.1%; 95% CI = (36.2%, 86.1%)]. The 95% CI of the replication effect size includes the original effect size for 12 replications [66.7%; 95% CI = (42.5%, 90.8%)]; if we also include the study in which the entire 95% CI exceeds the original effect size, this increases to 13 replications [72.2%; 95% CI = (49.3%, 95.1%)]. AER denotes the *American Economic Review* and QJE denotes the *Quarterly Journal of Economics*.

• Estimate
95% CI
Abeler et al., AER 2011 (33)
Ambrus and Greiner, AER 2012 (34)
Bartling et al., AER 2012 (35)
Charness and Dufwenberg, AER 2011 (36)
Chen and Chen, AER 2011 (37)
de Clippel et al., AER 2014 (38)
Duffy and Puzzello, AER 2014 (39)
Dulleck et al., AER 2011 (40)
Ericson and Fuster, QJE 2011 (41)
Fehr et al., AER 2013 (42)
Friedman and Oprea, AER 2012 (43)
Fudenberg et al., AER 2012 (44)
Huck et al., AER 2011 (45)
Ifcher and Zarghamee, AER 2011 (46)
Kessler and Roth, AER 2012 (47)
Kirchler et al, AER 2012 (48)
Kogan et al., AER 2011 (49)
Kuziemko et al., QJE 2014 (50)
-1
0
1
2
3

Figure: Camerer et al. (2016)

Table of Contents

1. Key Concepts and Definitions
2. Treatment Testing
3. Power Analysis
4. Clustering vs. Multilevel Modelling
5. Data Types
6. Meta-Regression

Dictator Game Experiment

Dictator game (with communication) example:

- Research question: *"Does the behavior of the dictator change if the receiver can communicate with them before the giving decision is made?"*
- Control: no communication.
- Treatment: with communication.
- Design: within-subject ($N = 30$ dictators) or between-subject ($N = 60$ dictators).

The dictator's endowment is \$100, and he gives \$$y$ to the receiver.

Tests of Normality

The independent-sample *t*-test is based on the assumption of normality of the two distributions whose means are being compared.

- We usually apply the Central Limit Theorem (CLT) when the samples being compared are sufficiently large.
- Central Limit Theorem (CLT): the (standardized) mean of a sufficiently large sample ($N \geq 30$) follows a standard normal distribution even when the sample is drawn from a sample that is not normal.

Tests for the normality of a population, based on sample data (H_0: the population is normally distributed):

- Skewness and kurtosis test.
- Shapiro-Wilk test.

Independent-Sample (Between Subject) Treatment Tests

Parametric treatment test: mean-comparison t-test

- H_0: $\mu_2 = \mu_1$.
- H_1: $\mu_2 > \mu_1$ (the prior belief predicts that communication has a positive effect on giving).
- Test statistic: $t = \frac{\bar{y_2} - \bar{y_1}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, which follows a $t(n_1 + n_2 - 2)$ distribution under H_0 (s_p is **pooled** standard deviation).
- Rejection rule (one-tailed test): reject H_0 if $t > t_{n_1+n_2-2,\alpha}$.

Independent-Sample (Between Subject) Treatment Tests

The bootstrap (Efron and Tibshirani, 1994): we can “bootstrap” the t-test without making any assumptions about the distribution of the data.

- 1) Apply t-test on the data set, obtaining a test statistic, $\hat{t}$.
- 2) Generate a number B of “bootstrap samples”. These are samples of the same size as the original sample. They are also drawn from the original sample **with replacement**. For each bootstrap sample, compute the test statistic, $\hat{t}_j^*$, $j = 1, ..., B$.
- 3) Compute the standard deviation s_B of the bootstrap test statistics, $\hat{t}_j^*$, $j = 1, ..., B$.
- 4) Obtain the new test-statistic $z_B = \frac{\hat{t}}{s_B}$.
- 5) Compare z_B against the standard normal distribution in order to find the “bootstrap p-value”.

Independent-Sample (Between Subject) Treatment Tests

The bootstrap:

- The number of bootstrap samples B should be chosen so that $\alpha(B+1)$ is a whole number (MacKinnon, 2002): B should be either 99 or 999 or 9999, as the size α is usually set to 0.01, 0.05, or 0.10.
- The bootstrap results in a different p-value each time the procedure is applied: fixing the results by choosing a large B or setting a random number seed.

Independent-Sample (Between Subject) Treatment Tests

Non-parametric treatment test: Mann–Whitney test (Wilcoxon rank-sum test)

- It does not rely on any strong distributional assumptions (e.g., normality).
- It depends on the assumption of equal variances between the two populations.

M-W test is based on a comparison of **the medians** of two samples (t-test is based on the comparison of two means).

- All of the observations from both samples are ranked by their value, with the highest rank being assigned to the largest value (ranks averaged in the event of a tie).
- The sum of ranks are found for each sample.
- The test is based on the comparison of these two sums.

M-W test is based solely on the ordinality of the data, and completely disregards the (possibly) rich cardinal information in the data.

Independent-Sample (Between Subject) Treatment Tests

Tests comparing entire distribution:

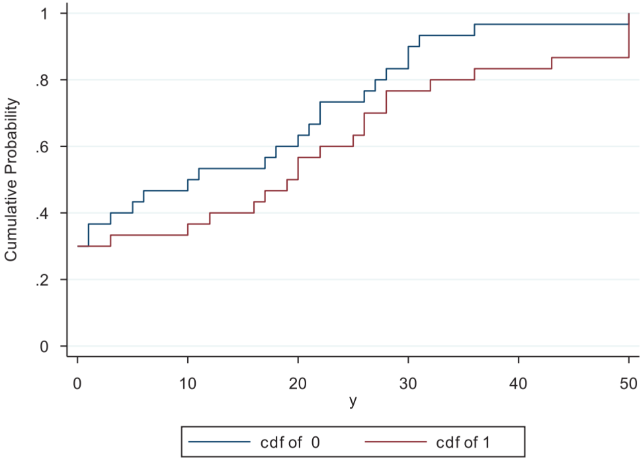
- The economic theory may not predict the precise nature of the treatment effect.
- What is shifted by the treatment? The mean of the distribution, the median, or the spread of the distribution?

We apply tests that are based on a comparison of the entire distributions under the two treatments, rather than a comparison of a particular functional (e.g., mean or median).

- Kolmogorov–Smirnov test: the test statistic is the maximum vertical distance between the two c.d.f.s and is used to judge whether the difference is significant.
- Epps-Singleton test: the test does not compare the two distributions directly, but instead compares the empirical characteristic functions.

Epps-Singleton test is applicable when the outcome has a discrete distribution (e.g., the number of questions answered correctly in a quiz).

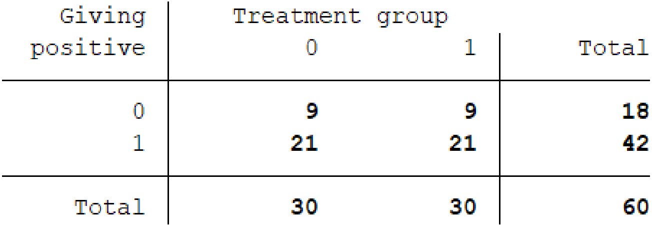
Independent-Sample (Between Subject) Treatment Tests



The cdf's of dictator's transfer under the no-communication treatment (higher line) and the communication treatment (lower line). The Kolmogorov–Smirnov test statistic is the maximum vertical distance between the two lines.

Independent-Sample (Between Subject) Treatment Tests

Binary outcomes (giving is positive or not): 2 X 2 cross-tabulation



```
   Giving |    Treatment group
 positive |         0          1 |     Total
----------+----------------------+----------
        0 |         9          9 |        18
        1 |        21         21 |        42
----------+----------------------+----------
    Total |        30         30 |        60
```

Independent-Sample (Between Subject) Treatment Tests

Binary outcomes (giving is positive or not): 2 X 2 cross-tabulation

- Pearson's Chi-squared test: the test compares the number in each cell with the number that would be expected if there were no treatment effect.
- Fisher's exact test: the test asks what is the probability of obtaining the combination of numbers in the tabulation, or a more extreme combination, for the given row totals and column totals, and this probability is the *p*-value for the test.

Fisher's exact test should be used when the Chi-squared approximation is likely to fail (e.g., the sample size is small, or when the numbers in some cells are very small).

Independent-Sample (Between Subject) Treatment Tests

Regression models:

- Treatment dummy is one of the explanatory variables.
- Outcome variable is the dependent variable.

The regression framework is the best approach to treatment testing:

- 1) It allows more treatment effects to be tested at the same time;
- 2) It allows explanatory variables other than the treatment variable to be controlled for (e.g., female-to-female giving tends to be lowest);
- 3) It allows adjustment for the dependence between observations (e.g., clustering).

Within Subject Treatment Tests

Within-subject tests are used when each subject is observed both with and without the treatment.

- Within-subject tests have greater statistical power: within-subject tests incorporate the additional information of the pairing of observations.
- The influence of confounding factors is reduced: each subject serves as his own control.
- Within-subject design has the concerns of “order effects”: the result of the test depends on the order of the two treatments (i.e., the experience of one treatment may impact on behavior in the following treatments).

Within Subject Treatment Tests

Parametric treatment test: paired mean-comparison *t*-test

- The test computes the difference in amount given between the two treatments for each subject, and then applies the *t*-test to test whether these differences have mean zero.

Within Subject Treatment Tests

Non-parametric treatment tests: Wilcoxon signed-ranks test

- The test is based on the differences in amount given between the two treatments, for each subject.
- The absolute differences are ranked from lowest to highest, so that the largest difference gets the highest value.
- These ranks are summed separately for the positive differences and the negative differences.
- The test is based on a comparison of these two rank sums.
- If there is no difference between the two treatments, these two rank sums should be roughly equal.

Wilcoxon signed-ranks test relies on the assumption that the distribution of paired differences is symmetric around the median.

Within Subject Treatment Tests

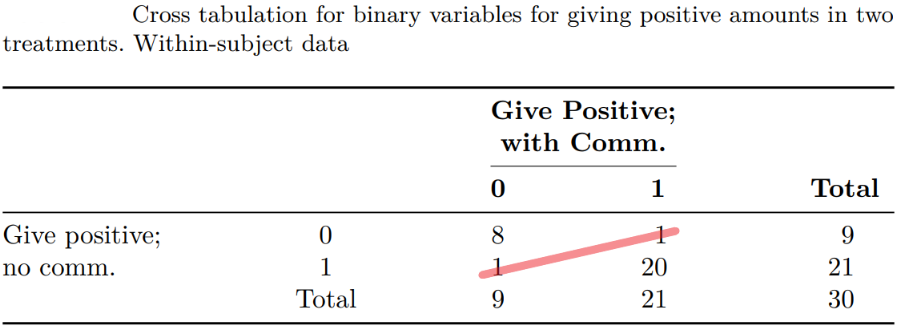
Non-parametric treatment tests: paired-sample sign test

- The test compares the number of positive differences to the number of negative differences, and asks if this difference is significantly different from one half according to a binomial distribution.
- It does not rely on the assumption that the distribution of paired differences is symmetric around the median.

Both non-parametric tests are based on the ordinality of the data, and disregard the cardinal information.

Within Subject Treatment Tests

Binary outcomes: McNemar's change test



Cross tabulation for binary variables for giving positive amounts in two treatments. Within-subject data

		Give Positive; with Comm.		
		0	**1**	**Total**
Give positive;	0	8	1	9
no comm.	1	1	20	21
	Total	9	21	30

- The test compares the two off-diagonal elements of cross-tabulation.
- If the number of subjects who change from not giving to giving as a result of communication is significantly greater than the number changing in the opposite direction, there is evidence of treatment effect.

Within Subject Treatment Tests

Regression models:

- It is imperative to allow adjustment for the dependence between observations (e.g., clustering).
- A paired sample can be viewed as a collection of clusters of size 2.

Parametric vs Non-parametric Treatment Tests

Parametric tests assume underlying statistical distributions in the data.

- For example, Student's t-test is reliable only if the sample is normal.

Non-parametric tests do not rely on any distribution.

- They can be applied even if parametric conditions of validity are not met.

For more treatment tests (more broadly, statistical tests):
https://help.xlstat.com/s/article/which-statistical-test-should-you-use?

Parametric vs Non-parametric Treatment Tests

Advantages of parametric tests

- 1) Parametric tests can provide reliable results with distributions that are skewed and non-normal;
- 2) Parametric tests can provide reliable results when the samples have different amounts of variability;
- 3) Parametric tests have greater statistical power.

Parametric vs Non-parametric Treatment Tests

Advantages of non-parametric tests

- 1) Non-parametric tests assess **the median** which can be better for some studies;
- 2) Non-parametric tests are valid when the sample size is small and the data are potentially non-normal;
- 3) Non-parametric tests can analyze ordinal data, ranked data, and outliers.

Generally, the parametric tests are more efficient, but the non-parametric tests are more robust.

Table of Contents

Power Analysis

The power π of a test is the probability of detecting an effect given that the effect really exists.

- 1) Power analysis is used to find the power of a test that has been performed;
- 2) … find the sample size required to perform a test with a given power.

Power Analysis

Let μ_1 and μ_2 be the population means of the control group and the treatment group respectively.

- H_0: $\mu_2 = \mu_1$ (i.e., the treatment has no effect).
- H_1: $\mu_2 - \mu_1 = \delta$ (i.e., the treatment has an effect with effect size δ).

For the problem of finding the required sample size to be properly defined, it is necessary to specify the value of δ:

- 1) From a pilot study;
- 2) From a previous study;
- 3) From prior beliefs;
- 4) Zhang and Ort mann (2013): the chosen effect size should be the smallest effect size with "economic significance".

Power Analysis, Independent Samples

To conduct the treatment test, first we need to decide:

- The size of the test α (e.g., $\alpha = 0.05$);
- The (adequate) power of the test π (e.g., $\pi = 0.80$ and $\beta = 1 - \pi = 0.20$).
- The sample sizes n_1 and n_2 (control and treatment respectively).

We usually use the independent samples t-test (let $n_1 = n_2 = n$ for simplicity):

- Test statistic: $t = \frac{\bar{y_2} - \bar{y_1}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{y_2} - \bar{y_1}}{s_p\sqrt{\frac{2}{n}}}$, following a $t(2n-2)$ distribution.
- Rejection rule (one-tailed test): reject H_0 if $t > t_{2n-2,\alpha}$, which becomes $t > z_\alpha$ as the value of n will be large enough.

Power Analysis, Independent Samples

The power of the test:

$$P(t > z_\alpha | \mu_2 - \mu_1 = \delta) = P(\frac{\bar{y_2} - \bar{y_1}}{s_p\sqrt{\frac{2}{n}}} > z_\alpha)|\mu_2 - \mu_1 = \delta) = \Phi(\frac{\delta - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}})$$

If the desired power of the test is $1 - \beta$, we have:

$$\Phi(\frac{\delta - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}}) = z_\beta$$

The required sample size:

$$n = \frac{2s_p^2(z_\alpha + z_\beta)^2}{\delta^2}$$

With $\alpha = 0.05$ and $\beta = 0.20$, the required sample size is $n = \frac{12.37s_p^2}{\delta^2}$.

Power Analysis Practice, Independent Samples

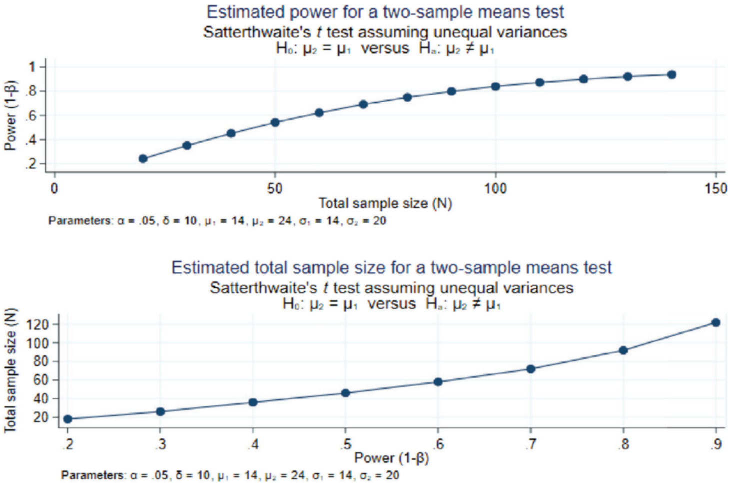
The dictator game experiment as an example:

- With sample size of 30 for each treatment, the power of the independent-sample t-test:

  $$P(t > z_{0.05} | \delta = 5.6) = \Phi(\frac{5.6 - 1.645 \times 15.9\sqrt{\frac{2}{30}}}{15.9\sqrt{\frac{2}{30}}}) = \Phi(-0.28) = 0.39.$$

- To bring the power up to 0.80, the sample size for each treatment has to be: $n = \frac{12.37 \times 15.9^2}{5.6^2} = 99.7$ (the required total sample size is about 200).

Power Analysis Practice, Independent Samples



Estimated power for a two-sample means test

Satterthwaite's *t* test assuming unequal variances

H_0: $\mu_2 = \mu_1$ versus H_a: $\mu_2 \neq \mu_1$

Parameters: $\alpha = .05$, $\delta = 10$, $\mu_1 = 14$, $\mu_2 = 24$, $\sigma_1 = 14$, $\sigma_2 = 20$

Estimated total sample size for a two-sample means test

Satterthwaite's *t* test assuming unequal variances

H_0: $\mu_2 = \mu_1$ versus H_a: $\mu_2 \neq \mu_1$

Parameters: $\alpha = .05$, $\delta = 10$, $\mu_1 = 14$, $\mu_2 = 24$, $\sigma_1 = 14$, $\sigma_2 = 20$

Upper panel: Power against total sample size; Lower Panel: Total required sample size against desired power.

Power Analysis Practice, Within Subject

In within-subject design, the **correlation** between the two treatments provide more information (e.g., if a subject gives generously in treatment 1, she is also likely to give generously in treatment 2):

- Paired-comparison tests have higher power than the corresponding independent-sample tests.
- Required total sample size is much lower.

Paired tests allow at least a 50% reduction in the sample size required for a given power, and the reduction can be much greater than 50% when the paired observations are highly correlated.

Power Analysis Practice, Within Subject

Order effects in within-subject design: the behaviors of subjects depend on the order of the two treatments.

- 1) Crossover designs

  Half of subjects see control followed by treatment, and the other half see treatment followed by control.

  Differences between these two groups would confirm the existence of an order effect, which need to be controlled for in treatment tests.

- 2) ABA designs

  All subjects experience three periods: control $\rightarrow$ treatment $\rightarrow$ control.

  ABA design can confirm the asymmetric effects between (control $\rightarrow$ treatment) and (treatment $\rightarrow$ control) within the same subject.

Monte Carlo Simulations of Tests

Which treatment test performs better (i.e., correct size α and adequate power π)?

- 1) Normally distributed data.
- 2) Non-normal or highly skewed data.

We use a Monte Carlo simulation to find:

- The actual size α of each test when there is no effect $\delta = 0$ and nominal $\alpha = 0.05$.
- The power π of each test when the true effect size is $\delta = 0.5$ and nominal $\alpha = 0.05$.

Monte Carlo Simulations of Tests

We assume the data generating process:

$$
\begin{aligned}
& y_i = 10 + \delta d_i + \varepsilon_i \\
& d_i = 0 \text{ if } i \leq \frac{n}{2} \text{ and } d_i = 1 \text{ if } i > \frac{n}{2} \\
& E[\varepsilon_i] = 0 \text{ and } V(\varepsilon_i) = 1
\end{aligned}
$$

- y_i: outcome variable;
- d_i: dummy variable of treatment;
- δ: treatment effect;
- n: total sample size.

Monte Carlo Simulations of Tests

Monte Carlo estimates of size and power of four tests: *t*-test; Mann–Whitney (MW); Kolmogorov–Smirnov (KS); Epps–Singleton (ES). All tests have nominal size 0.05. Data generating process in (3.9). Three different distributions assumed for the error term in (3.9): normal; uniform; skew ($\chi^2(3)$). 50 observations per treatment

	Normal		**Uniform**		**Skew** ($\chi^2(3)$)	
	Size	**Power**	**Size**	**Power**	**Size**	**Power**
t-test	0.051	0.692	0.051	0.570	0.051	0.710
MW	0.050	0.672	0.051	0.534	0.052	0.880
KS	~~0.041~~	0.533	~~0.040~~	0.303	~~0.039~~	0.880
ES	0.054	0.488	~~0.065~~	0.782	0.053	0.970

The distribution (e.g., non-normality, asymmetry) of the data is very important in the choice of treatment test.

No. of Subjects and Tasks

Many experiments require subjects to do multiple tasks, we need to consider the number of tasks (T):

- We may face a trade-off between the number of subjects n and the number of tasks T due to budget constraint.

Given an increase in the budget, is it more beneficial to increase n, or T?

No. of Subjects and Tasks

Assuming dependence at the subject and group levels, the 3-level data generating process:

$$y_{ijt} = \beta' x_{it} + \delta d_i + u_i + v_j + \varepsilon_{ijt}$$
$$V(u_i) = \sigma_u^2,\ V(v_i) = \sigma_v^2,\ V(\varepsilon_{ijt}) = \varepsilon_u^2$$
$$i = 1, ..., n,\ j = 1, ..., J,\ t = 1, ..., T$$

- y_{ijt}: decision made by subject i in group j in task t;
- x_{it}: a vector of control variables;
- d_i: dummy variable indicating treatment;
- δ: treatment effect;
- u_i: subject-specific random effect;
- v_i: group-specific random effect;
- ε_{ijt}: observation-specific error term.

No. of Subjects and Tasks

Monte Carlo results using (3.10) as DGP. Power of between-subject test (with true effect size 0.50) and within-subject test (with true effect size 0.05), at different combinations of number of subjects (n) and number of tasks (T)

	Between-Subject Test ($\boldsymbol{\delta = 0.50}$)			**Within-Subject Test** ($\boldsymbol{\delta = 0.05}$)		
	$\boldsymbol{T = 50}$	$\boldsymbol{T = 100}$	$\boldsymbol{T = 150}$	$\boldsymbol{T = 50}$	$\boldsymbol{T = 100}$	$\boldsymbol{T = 150}$
$n = 40$	0.24	0.26	~~0.28~~	0.20	0.47	~~0.75~~
$n = 80$	0.25	~~0.34~~	0.35	0.44	~~0.71~~	0.91
$n = 120$	~~0.39~~	0.38	0.35	~~0.67~~	0.81	0.97

In between-subject design, increases in n might be more beneficial; in within-subject design, increases in T might be more beneficial.

Table of Contents

1. Key Concepts and Definitions
2. Treatment Testing
3. Power Analysis
4. Clustering vs. Multilevel Modelling
5. Data Types
6. Meta-Regression

Clustering vs. Multilevel Modelling

The regression approach allows adjustment for the dependence between observations:

- At the level of the individual subject (e.g., paired data or multiple tasks);
- ... the group of subjects that are interacting with each other (e.g., the public goods games);
- ... the experimental session;
- ... the university, or city, or region, etc.

Clustering vs. Multilevel Modelling

There are two common methods of adjusting for dependence: clustering and multilevel modelling.

- Clustering is the process of obtaining “cluster-robust” standard errors for OLS estimates.
- Multilevel modelling fully respects the clustered structure in estimation, and hence results in an efficient estimator of the model parameters, as well as unbiased standard errors.

The default OLS standard errors that ignore clustering can greatly underestimate the true OLS standard errors.

Clustering vs. Multilevel Modelling

Assuming dependence at the subject and group levels, the 3-level data generating process:

$$\begin{aligned}
&y_{ijt} = \beta' x_{it} + \delta d_i + u_i + v_j + \varepsilon_{ijt} \\
&V(u_i) = \sigma_u^2,\ V(v_i) = \sigma_v^2,\ V(\varepsilon_{ijt}) = \varepsilon_u^2 \\
&i = 1, ..., n,\ j = 1, ..., J,\ t = 1, ..., T
\end{aligned}$$

- y_{ijt}: decision made by subject i in group j in task t;
- x_{it}: a vector of control variables;
- d_i: dummy variable indicating treatment;
- δ: treatment effect;
- u_i: subject-specific random effect;
- v_i: group-specific random effect;
- ε_{ijt}: observation-specific error term.

Clustering vs. Multilevel Modelling

Test clustering vs. multilevel modelling with Monte Carlo method:

- Which tests are correctly sized?
- Of those which are correctly sized, which has highest power?

Clustering vs. Multilevel Modelling

Results from Monte Carlo experiment with DGP (3.10). All tests have nominal size 0.05

	Between-Subject Test		**Within-Subject Test**	
	Size	**Power ($\boldsymbol{\delta = 0.5}$)**	**Size**	**Power ($\boldsymbol{\delta = 0.5}$)**
OLS no clustering ✖	~~0.46~~	0.68	~~0.02~~	0.07
OLS with clustering at subject level	~~0.15~~	0.41	0.09	0.31
OLS with clustering at group level	0.07	0.25	0.09	0.33
Random effects no clustering	~~0.13~~	0.41	0.05	0.31
Random effects with clustering at subject level	~~0.15~~	0.41	0.09	0.31
Random effects with clustering at group level	0.07	0.25	0.08	0.33
Multi-level model	0.06	0.27	0.05	0.31

Clustering vs. Multilevel Modelling

Recommendations about clustering vs. multilevel modelling:

- 1) OLS without clustering is always avoided.
- 2) Multilevel modelling might be preferred.
- 3) ? If clustering is to be used, it is preferable to be clustered at the highest possible level (e.g., group rather than subject).

Cameron and Miller (2015): *"... there is no clear-cut definition of "few" (clusters). Depending on the situation "few" may range from less than 20 clusters to less than 50 clusters in the balanced case, and even more clusters in the unbalanced case."*

Table of Contents

Properties of Estimators

Consistency of estimators

- An estimator T_n of parameter θ is consistent, if it converges in probability to the true value of the parameter: $\lim_{n\to\infty} \Pr\left(|T_n - \theta| > \varepsilon\right) = 0$.
- The distributions of the estimates become more and more concentrated near the true value of the parameter, so that the probability of the estimator being arbitrarily close to θ converges to one.

Unbiasedness of estimators

- An estimator T of parameter θ on observable data x is unbiased if the expected value of the estimator matches that of the parameter: $E_{x|\theta}[T] = \theta$.
- Unbiasedness has nothing to do with the number of observation used in the estimation.

Properties of Estimators

Efficiency of estimators

- The efficiency of an estimator T of a parameter θ: $e(T) = \frac{1}{\mathrm{var}(T)\mathcal{I}(\theta)}$.
- The efficiency is the precision of the estimator divided by the upper bound of the precision (the Fisher information of the sample $\mathcal{I}(\theta)$).
- An efficient estimator is characterized by a small variance, indicating that there is a small deviance between the estimated value and the true value.
- A more efficient estimator needs fewer observations than a less efficient one to achieve a given performance.

Binary Data

Examples

- 1) Decision under risk: binary choices between a safe lottery and a risky lottery.
- 2) Intertemporal choice: binary choices between an earlier smaller option and a later larger option.
- 3) Ultimatum game: binary decision of the receiver to reject or accept the offer.

Model and estimator

- 1) Binary probit (MLE): $P(y_i = 1|X_i) = \Phi(X_i^T\beta)$, where Φ is the c.d.f. of standard normal distribution.
- 2) Binary logit (MLE): $P(y_i = 1|X_i) = \Lambda(X_i^T\beta)$, where Λ is the c.d.f. of logistic distribution.

Ordinal Data

Examples

- 1) Strength of preference (Butler et al., 2014) (stated after binary choice between two lotteries): “completely unsure”, “fairly unsure”, “neither unsure nor sure”, “fairly sure”, “completely sure”.
- 2) Strength of agreement (Likert scale in surveys): “strongly disagree”, “disagree”, “neither agree nor disagree”, “agree”, “strongly agree”.

Ordered strength of preference vs. binary choice:

- Models using ordered strength of preference are more efficient since ordinal data embodies more information than binary data.
- The stated strength of preference cannot be incentive compatible, but binary choice between the two lotteries can be incentive compatible.

Ordinal Data

Model and estimator

- 1) Ordered probit (MLE).
- 2) Ordered logit (MLE).
- They assume an underlying unobserved latent variable y_i^* as the dependent variable in a linear regression model: $y_i^* = x_i'\beta + \varepsilon_i$, where ε_i is assumed to be standard normal or logistic.
- The relationship between the latent variable y_i^* and the observed variable y is: $y = j - 1$ if $\kappa_{j-1} < y^* < \kappa_j$, where the parameters κ_j, $j = 1, ..., J - 1$, are the "cut-point" parameters.

Interval Data

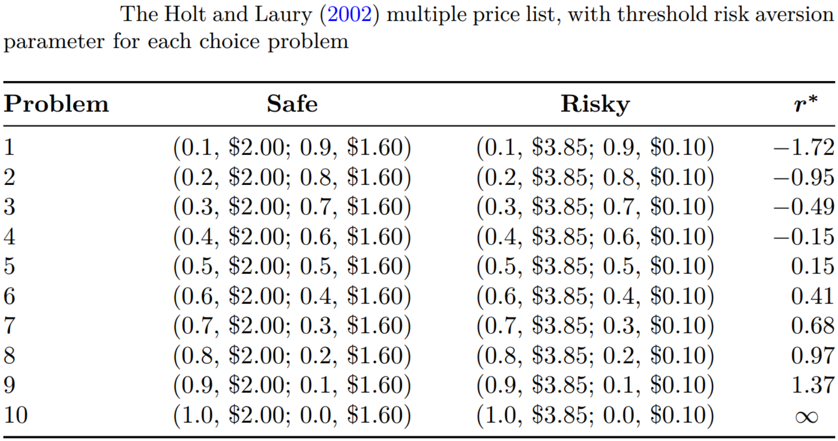
Examples

- 1) Risk preference elicitation with multiple price list.
- 2) Time preference elicitation with multiple price list.
- 3) Income groups in surveys.

Model and estimator

- Interval regression (MLE).
- Applying ordered probit model to interval data: the estimator is inefficient (because it involves the unnecessary estimation of the known cut-points).
- Applying OLS to the midpoints of the intervals: the estimator is inconsistent (Stewart, 1983).

Interval Data



The Holt and Laury (2002) multiple price list, with threshold risk aversion parameter for each choice problem

Problem	Safe	Risky	r^*
1	(0.1, \$2.00; 0.9, \$1.60)	(0.1, \$3.85; 0.9, \$0.10)	-1.72
2	(0.2, \$2.00; 0.8, \$1.60)	(0.2, \$3.85; 0.8, \$0.10)	-0.95
3	(0.3, \$2.00; 0.7, \$1.60)	(0.3, \$3.85; 0.7, \$0.10)	-0.49
4	(0.4, \$2.00; 0.6, \$1.60)	(0.4, \$3.85; 0.6, \$0.10)	-0.15
5	(0.5, \$2.00; 0.5, \$1.60)	(0.5, \$3.85; 0.5, \$0.10)	0.15
6	(0.6, \$2.00; 0.4, \$1.60)	(0.6, \$3.85; 0.4, \$0.10)	0.41
7	(0.7, \$2.00; 0.3, \$1.60)	(0.7, \$3.85; 0.3, \$0.10)	0.68
8	(0.8, \$2.00; 0.2, \$1.60)	(0.8, \$3.85; 0.2, \$0.10)	0.97
9	(0.9, \$2.00; 0.1, \$1.60)	(0.9, \$3.85; 0.1, \$0.10)	1.37
10	(1.0, \$2.00; 0.0, \$1.60)	(1.0, \$3.85; 0.0, \$0.10)	∞

If a subject chooses S on problems 1–3, and chooses R on problems 4–10, their risk attitude (i.e., CRRA coefficient r) is between -0.49 and -0.15.

Censored Data

Examples

- 1) Dictator game, ultimatum game, trust game: the transfer from the first mover (or the second mover).
- 2) Public goods game: the contribution to a public fund.

Model and estimator

- 1) Tobit model (MLE).
- 2) Hurdle model (MLE).
- The lower limit is to the subject's transfer or contribution, usually zero.
- The upper limit is his endowment.
- When OLS is applied to censored data, the slope estimates tend to be seriously biased towards zero.

Censored Data

In a public goods game:

- Subjects may first decide whether they want to be a free rider (extensive margin);
- Of those who do not want to be a free rider, they then decide how much to contribute (intensive margin).

Two distinct types of zero observations: censored zeros, as in the tobit model, and "zero types".

- A **zero type** is a subject who always contribute or transfer zero whatever the circumstances.
- In public goods games, a zero type is a "free rider", and in dictator games, they are "selfish".

Censored Data

Hurdle models provide separate equations for the bounded and the unbounded outcomes (tobit models use the same equation for both).

- Hurdle models assume the unbounded outcomes are the result of clearing a hurdle.
- When the hurdle is not cleared, bounded outcomes result.
- The first equation determines whether we clear the hurdle (i.e., to be a contributor or not).
- The second determines the value of the outcome conditional on having cleared the hurdle (i.e., if being a contributor, how much to contribute).

Censored Data

Engel and Moffatt (2012) use a hurdle model to test “house money effect” (i.e., people tend to take on increased risk subsequent to a successful investment experience) in the context of a public goods game.

- House money increases the probability of passing the hurdle, i.e., being a “potential contributor”.
- House money can change a subject from one type to another.

Continuous (Exact) Data

Example

- 1) Stated certainty equivalent of a risky lottery.
- 2) WTP and WTA elicited from BDM mechanism.

Model and estimator

- Linear regression (OLS).
- Greater estimation efficiency relative to discrete data: the exact value embodies more information than interval or ordinal data.
- Underestimation of the degree of risk aversion: when subjects are asked to value a lottery, there is a tendency for the response to be biased towards the expected value of the lottery.
- BDM mechanism is too complicated for subjects to understand: this will divert attention from the valuation task itself.

Table of Contents

1. Key Concepts and Definitions
2. Treatment Testing
3. Power Analysis
4. Clustering vs. Multilevel Modelling
5. Data Types
6. Meta-Regression

Meta-Analysis Definitions

Meta-analysis: statistical analysis that combines the results of multiple scientific studies addressing the same question.

- Multiple scientific studies address the same question, but each individual study may have some degree of error.
- The aim is to derive a pooled estimate closest to the unknown common truth based on how this error is perceived.
- It can contrast results from different studies and identify patterns among study results or study the sources of disagreement among those results.

Meta-Analysis Definitions

Meta-regression: regression analysis to combine, compare, and synthesize research findings from multiple studies while adjusting for the effects of covariates on a response variable.

- The dependent variable is a summary statistic, perhaps a regression parameter, drawn from each study.
- The independent variables may include characteristics of the method, design and data used in these studies.

It can identify the extent which the particular choice of methods, design and data affect reported results.

Meta-Regression Steps

Steps of meta-regression analysis (Stanley, 2001):

- 1) Include all relevant studies from a standard database;

  All studies, published or not (to reduce potential biases introduced by any nonrandom selection of studies).

- 2) Choose a summary statistic and reduce the evidence to a common metric;

  Regression coefficients, elasticities, t-values etc (some transformation or standardization of reported statistics might be required).

- 3) Choose moderator variables;

  Study characteristics that are thought to be consequential (e.g., dummy variables for different data sets and econometric choices).

Meta-Regression Steps

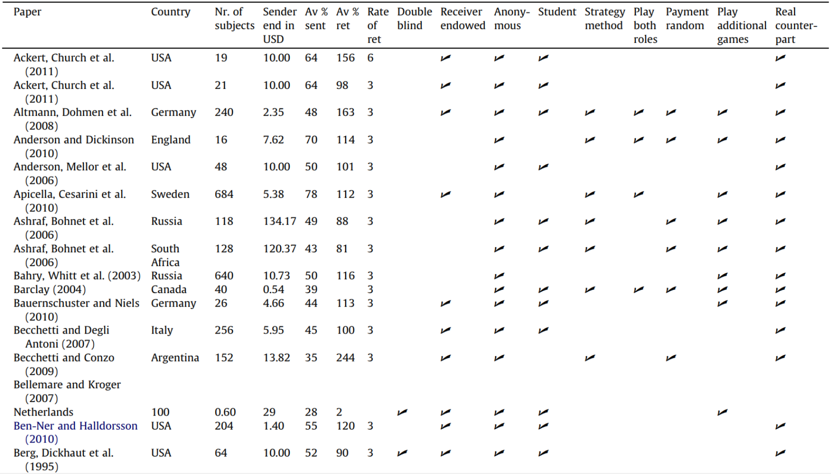
- 4) Conduct a meta-regression analysis;

  Study-to-study variation to be explained in an empirical literature (if one variable is found to be important, then empirical studies in this area should include this variable).
- 5) Subject the meta-regression analysis to specification testing.

  Autocorrelation, heteroskedasticity, misspecification etc.

Meta-Regression Example

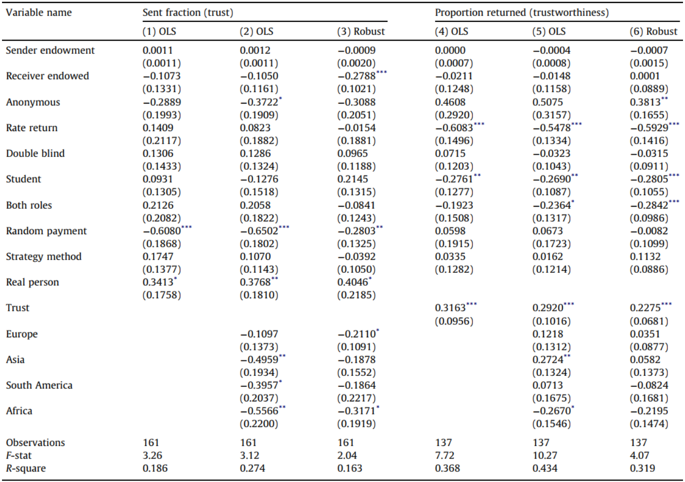
Meta-regression analysis of trust games (Johnson and Mislin, 2011): selected papers X selected factors.



Paper	Country	Nr. of subjects	Sender end in USD	Av % sent	Av % ret	Rate of ret	Double blind	Receiver endowed	Anony-mous	Student	Strategy method	Play both roles	Payment random	Play additional games	Real counter-part
Ackert, Church et al. (2011)	USA	19	10.00	64	156	6		✓	✓	✓					✓
Ackert, Church et al. (2011)	USA	21	10.00	64	98	3		✓	✓	✓					✓
Altmann, Dohmen et al. (2008)	Germany	240	2.35	48	163	3		✓	✓	✓	✓	✓	✓	✓	✓
Anderson and Dickinson (2010)	England	16	7.62	70	114	3			✓		✓	✓	✓	✓	✓
Anderson, Mellor et al. (2006)	USA	48	10.00	50	101	3			✓	✓					✓
Apicella, Cesarini et al. (2010)	Sweden	684	5.38	78	112	3		✓	✓		✓	✓		✓	✓
Ashraf, Bohnet et al. (2006)	Russia	118	134.17	49	88	3			✓	✓	✓		✓	✓	✓
Ashraf, Bohnet et al. (2006)	South Africa	128	120.37	43	81	3			✓	✓	✓		✓	✓	✓
Bahry, Whitt et al. (2003)	Russia	640	10.73	50	116	3			✓					✓	✓
Barclay (2004)	Canada	40	0.54	39		3			✓	✓	✓	✓	✓	✓	✓
Bauernschuster and Niels (2010)	Germany	26	4.66	44	113	3		✓	✓	✓				✓	✓
Becchetti and Degli Antoni (2007)	Italy	256	5.95	45	100	3		✓	✓	✓					✓
Becchetti and Conzo (2009)	Argentina	152	13.82	35	244	3		✓	✓		✓		✓		✓
Bellemare and Kroger (2007)															
Netherlands	100	0.60	29	28	2		✓	✓	✓	✓				✓	
Ben-Ner and Halldorsson (2010)	USA	204	1.40	55	120	3		✓	✓	✓					✓
Berg, Dickhaut et al. (1995)	USA	64	10.00	52	90	3	✓	✓	✓	✓					✓

Meta-Regression Example

Meta-regression analysis of trust games (Johnson and Mislin, 2011): regressing “trust” and “trustworthiness” on selected factors.



Variable name	Sent fraction (trust)			Proportion returned (trustworthiness)		
	(1) OLS	(2) OLS	(3) Robust	(4) OLS	(5) OLS	(6) Robust
Sender endowment	0.0011	0.0012	−0.0009	0.0000	−0.0004	−0.0007
	(0.0011)	(0.0011)	(0.0020)	(0.0007)	(0.0008)	(0.0015)
Receiver endowed	−0.1073	−0.1050	−0.2788***	−0.0211	−0.0148	0.0001
	(0.1331)	(0.1161)	(0.1021)	(0.1248)	(0.1158)	(0.0889)
Anonymous	−0.2889	−0.3722*	−0.3088	0.4608	0.5075	0.3813**
	(0.1993)	(0.1909)	(0.2051)	(0.2920)	(0.3157)	(0.1655)
Rate return	0.1409	0.0823	−0.0154	−0.6083***	−0.5478***	−0.5929***
	(0.2117)	(0.1882)	(0.1881)	(0.1496)	(0.1334)	(0.1416)
Double blind	0.1306	0.1286	0.0965	0.0715	−0.0323	−0.0315
	(0.1433)	(0.1324)	(0.1188)	(0.1203)	(0.1043)	(0.0911)
Student	0.0931	−0.1276	0.2145	−0.2761**	−0.2690**	−0.2805***
	(0.1305)	(0.1518)	(0.1315)	(0.1277)	(0.1087)	(0.1055)
Both roles	0.2126	0.2058	−0.0841	−0.1923	−0.2364*	−0.2842***
	(0.2082)	(0.1822)	(0.1243)	(0.1508)	(0.1317)	(0.0986)
Random payment	−0.6080***	−0.6502***	−0.2803**	0.0598	0.0673	−0.0082
	(0.1868)	(0.1802)	(0.1325)	(0.1915)	(0.1723)	(0.1099)
Strategy method	0.1747	0.1070	−0.0392	0.0335	0.0162	0.1132
	(0.1377)	(0.1143)	(0.1050)	(0.1282)	(0.1214)	(0.0886)
Real person	0.3413*	0.3768**	0.4046*			
	(0.1758)	(0.1810)	(0.2185)			
Trust				0.3163***	0.2920***	0.2275***
				(0.0956)	(0.1016)	(0.0681)
Europe		−0.1097	−0.2110*		0.1218	0.0351
		(0.1373)	(0.1091)		(0.1312)	(0.0877)
Asia		−0.4959**	−0.1878		0.2724**	0.0582
		(0.1934)	(0.1552)		(0.1324)	(0.1373)
South America		−0.3957*	−0.1864		0.0713	−0.0824
		(0.2037)	(0.2217)		(0.1675)	(0.1681)
Africa		−0.5566**	−0.3171*		−0.2670*	−0.2195
		(0.2200)	(0.1919)		(0.1546)	(0.1474)
Observations	161	161	161	137	137	137
F-stat	3.26	3.12	2.04	7.72	10.27	4.07
R-square	0.186	0.274	0.163	0.368	0.434	0.319

Meta-Regression Example

Meta-regression analyses in experimental economics:

- Trust game: Johnson and Mislin (2011)
- Dictator game: Engel (2011)
- Ultimatum game: Cochard et al. (2021)
- Public good game: Zelmer (2003)
- Quasi-hyperbolic discounting: Cheung et al. (2021)
- Cognitive ability and risk aversion: Lilleholt (2019)

References I

- Peter G Moffatt. “Experimetrics: A Survey”. In: *Foundations and Trends in Econometrics* 11.1-2 (2021), pp. 1–152.
- Tom D Stanley. “Wheat from chaff: Meta-analysis as quantitative literature review”. In: *Journal of Economic Perspectives* 15.3 (2001), pp. 131–150.