Experimental Methodology

-Definition of an experiment

-Lab and field experiments

-Stakes and deception

-Hypothesis testing, p-value and statistical power

-P-hacking and forking

-Preregistration

Uses of the Term ”Experiment” in Economics

-A primary data collection to estimate the causal effect on behavior of varying one (or several) variables (e.g. a test of if the stakes ($10 versus $20) affects behavior in the ultimatum game:

Between subjects treatment comparisons

Within subjects treatment comparisons

-A primary data collection to test if behavior in for instance a specific game is consistent with a theoretical prediction (e.g. a test of if behavior in the ultimatum game is consistent with theoretical predictions).

-A primary data collection of the difference in behavior between different groups, e.g. testing if behavior differ between women and men.

Classical meaning of ”Experiment”

Between subjects treatment comparison with randomization to treatments (RCT-Randomized Controlled Trial); randomization key to identify causal effects.

Experimental Terminology

-Subjects (the participants in an experiment)

-Treatments (the experimental groups/conditions in a study; at least one variable varies between each treatment)

-Sessions (the number of separate data collection ”events”)

-Rounds (the number of times a subject ”repeats” a ”task” in an experiment)

-Stranger and partner matching (stranger matching implies matching with new subjects in each round and partner matching implies matching with the same players in each round)

-Single-blind(decisions of subjects anonymous with respect to other subjects) and double-blind (decisions also anonymous with respect to the experimenter) designs (term used differently than in other fields where it typically means blinded to the treatment they receive)

-Direct response method versus strategy method (e.g. responding to the actual offer in an ultimatum game versus responding to all possible offers prior to knowing the actual offer)

Laboratory Experiments and Field Experiments

The distinction between lab and field experiments is not obvious; but a field experiment typically means that the experiment is carried out under more realistic circumstances.

Harrison and List (JEL 2004) taxonomy:

-Conventional lab experiment (a subject pool of students, abstract framing and an imposed set of rules).

-Artefactual field experiment (the same as a conventional lab experiment, but with a nonstandard subject pool).

-Framed field experiment (the same as an artefactual field experiment, but with field context in the commodity, task, or information set).

-Natural field experiment (the same as a framed field experiment, but where the environment is one where the subjects naturally undertakes these tasks and where the subjects do not know that they are in an experiment).

Online Experiments

Many experiments are carried out online; lab or field experiments?

Advantages and disadvantages?

Amazon Mechanical Turk (AMT): Online labor market for small ”jobs” often used for experiments.

Different Ways of Executing a Lab Experiment used in the Literature

10 sessions assumed in example, but the number of sessions can vary.

1. Subjects can decide which out of ten sessions (time and day) to sign-up for. The experimenters run 5 sessions of treatment A followed by 5 sessions of treatment B. Problems?

2. Subjects can decide which out of ten sessions (time and day) to sign-up for. The experimenters randomly allocates treatment A to five sessions and treatment B to five sessions. Problems?

3. Subjects can sign up for the experiment, but not a specific session (time and day). The experimenters randomly allocates signed-up subjects to five sessions (time and day) with treatment A and five sessions with treatment B. Problems?

4. Subjects can decide which out of five ”parallel-sessions” (time and day) to sign-up for. When subjects show-up the experimenters randomly allocates subjects to two rooms and carry out treatment A in one room and treatment B in the other room and runs both sessions at the same time. Problems?

5. Subjects can decide which out of ten sessions (time and day) to sign-up for. When subjects are seated in the lab they are in each session randomly allocated to treatment A or B that are carried out in the same room. Problems?

Deception in Experiments

A strong norm in experimental economics that deception, lying to the subjects, is not allowed (whereas deception is relatively common in psychology experiments).

Economics journals typically do not publish experiments based on deception (although somewhat different norms in different sub-fields; e.g. the use of fictitious applications/resumes in labor economics).

Advantages and disadvantages of using deception in experiments?

Stakes in Experiments

Strong empasis on real stakes in experiments in economics; a difference to experiments in psychology where hypothetical decisions are relatively common.

Typically a show-up fee and a ”performance” based payment based on a reasonable hourly wage-rate.

In experiments with multiple rounds or decisions one round or decision often selected for real payment.

Low stakes in AMT studies.

The importance of stakes can vary depending on the type of decision, and stakes can also affect the variance of decisions. Important in for instance ”social desirability tasks”.

Hypotheses Testing and Type 1 and Type II errors

Null hypothesis H0: e.g. stakes have no effect on proposals in the ultimatum game.

Alternative Hypothesis H1: e.g. stakes have an effect on proposals in the ultimatum game.

Type 1 error: Rejecting the null hypothesis if it is true (false positive).

Type II error: Failing to reject the null hypothesis when the alternative hypothesis is true (false negative).

Which error is most important?

P-value, Power and Type 1 and Type II errors

p-value: the probability of observing an effect size that is equal to or higher than the observed effect size if the null hypothesis is true. A p-value threshold of 0.05 often used for ”statistical significance”. The probability of making a type 1 Error.

Statistical power: the probability of observing an hypothesized effect size if the hypothesized effect size is the true effect size. 1-power the probability of making a type II error.

For a given sample size there is a trade-off between type I and type II errors (e.g. using a lower significance threshold increases the type II error probability for a given sample size).

Interpretation of Statistically Significant Findings

p-values and p-value thresholds often misinterpreted. A common misconception that a significant finding with a p-value threshold of 0.05, implies 95% chance that the alternative hypothesis is correct.

Pre-experimental odds of a tested hypothesis being true if ”statistically significant finding”= $(\pi1 * \text{Ϙ})/(\pi0 * \alpha)$

$\pi1$=prior of hypothesis H1 being correct

$\pi0$=prior of null hypothesis H0 being correct

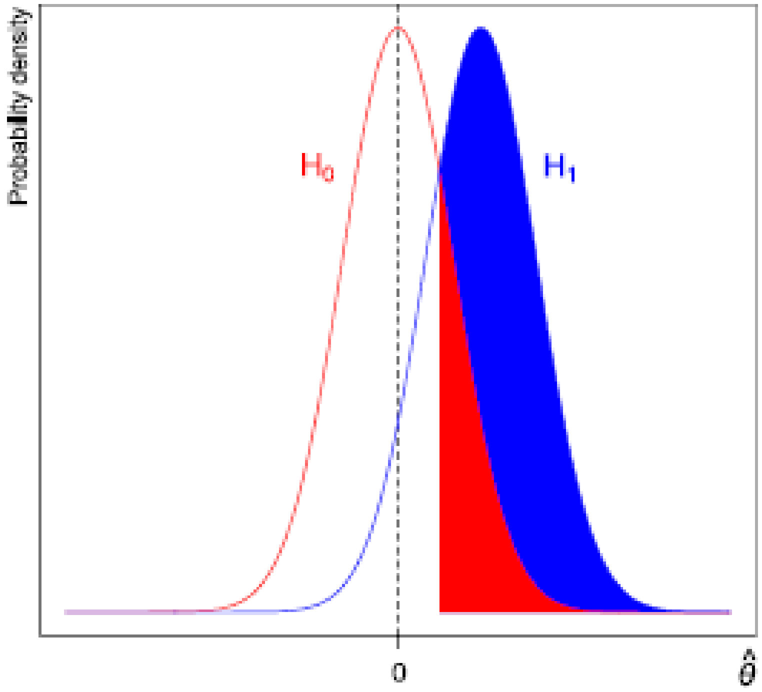
α=statistical significance threshold (type 1 error)

Ϙ=statistical power (1-type II error)

Rearranging gives: $(\pi1/\pi0) * (\text{Ϙ}/\alpha)$

e.g. prior of hypothesis being true=0.2, power=0.8, p-value threshold=0.05 gives odds of 4 (and the pre-experimental rejection ratio of H1 to H0 $(\text{Ϙ}/\alpha)$ is 16).

The risk of a ”significant” finding being a false positive increases with low power, low priors and higher p-value threshold.



The red curve shows the probabilities of observing different mean effect sizes if the null hypothesis (H0) of a zero effect is true given a specific sample size; and the blue curve shows the probabilities of observing different mean effect sizes if the alternative hypothesis (H1 is true). The red filled area is the probability of observing a significant result if the null hypothesis is true (for a one-sided test with α=0.05). The filled blue area plus the filled red area is the probability of observing a significant result if H1 is true, and equals the statistical power. The ratio of (blue+red)/red is equal to Ϙ/α on the previous slide; the pre-experimental rejection ratio of H1 to H0.

Figure from Bayarri et al. Journal of Mathemathical Psychology 2016.

How the false positive rate (the fraction of statistically significant findings that are false positives) depends on power, prior odds and the p-value threshold.

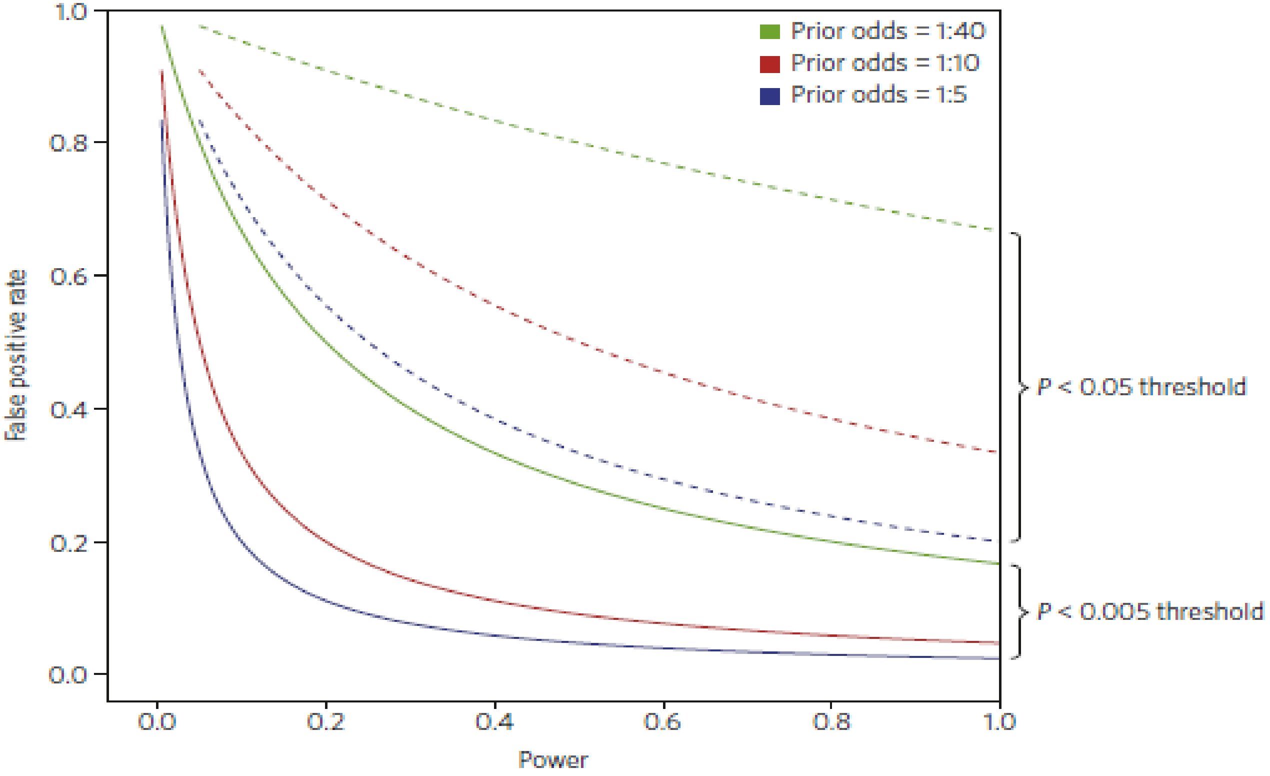


Fig. 2 | Relationship between the *P* value threshold, power, and the false positive rate. Calculated according to equation (2), with prior odds defined as $\frac{1-\phi}{\phi} = \frac{\Pr(H_1)}{\Pr(H_0)}$. For more details, see the Supplementary Information.

Post Experimental Rejection Odds

Once the experiment has been conducted the p-value is compared to the statistical significance threshold to see if the result is ”statistically significant” or not.

But the exact p-value also carries further information; the lower is the p-value for a ”statistically significant finding” the more strong is the support for the tested hyppothesis H1.

Post-experimental odds of a tested hypothesis being true= $(\pi1/\pi0) * BF$

BF is the post-experimental rejection ratio often called the Bayes Factor (BF). It is the probability of the observed data given H1 divided by the probability of the observed data given H0 (BF is the factor that new evidence updates the prior odds with). To estimate BF it is necessary to make some assumption about the prior for H1.

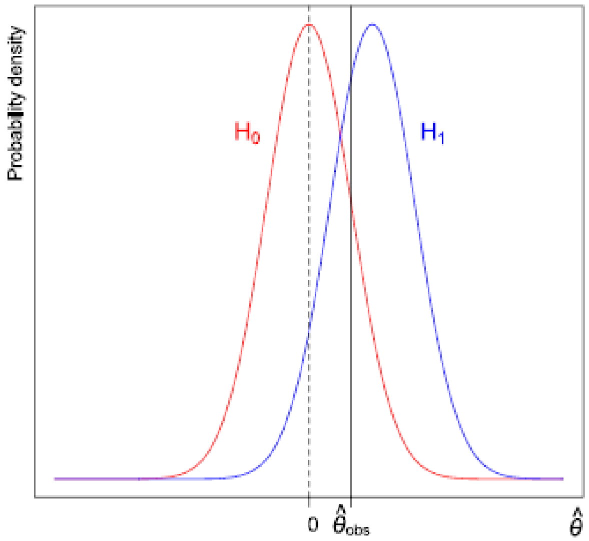
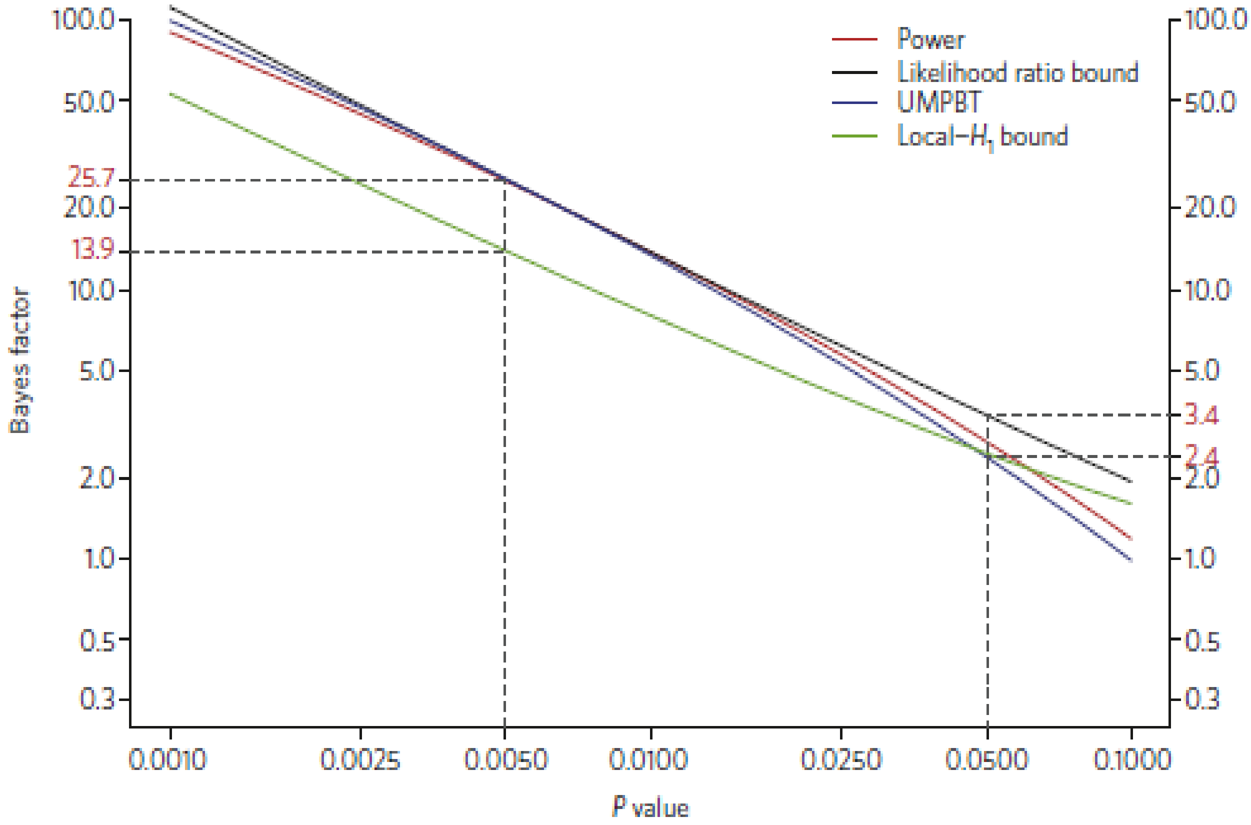


Fig. 2. The post-experimental rejection ratio (Bayes factor) is the ratio of the probability density of $\hat{\theta}_{obs}$ under H_1 to the probability density of $\hat{\theta}_{obs}$ under H_0.

The same graph as before, but with the observed mean effect size added to the graph. A point prior assumed equal to the hypothesized effect size of H1. The Bayes Factor (BF) is the ratio of the probability of observing the data for H1 to the probability of observing the data for H0. The problem in estimating the BF is that it's not obvious how to define the prior for H1 (some distribution over possible effect sizes most realistic rather than a point prior as in the graph).

Figure from Bayarri et al. Journal of Mathemathical Psychology 2016.



The relationship between the p-value and the Bayes Factor for different ways of estimating the Bayes Factor.

Figure from Benjamin et al. Nature Human Behaviour 2018

Statistical Power and Sample Size of Experiments

To estimate the statistical power of a study information about the following is needed: hypothesized effect size, standard deviation, sample size, and p-value threshold of significance test.

To estimate the sample size needed to have for instance 90% statistical power information about the following is needed: hypothesized effect size, standard deviation, and p-value threshold of significance test.

Statistical power given a specific sample size or the sample size needed for a specific power can be estimated using for instance different online power calculators.

What is the statistical power of a replication of an original study that found an effect with a p-value of 0.049 if the hypothesized effect size is the same as found in the original study and the replication has the same sample size as the original study (and uses a $p<0.05$ p-value threshold)?

Minimum detectable effect size (MDE): The MDE to have 80% power to detect an effect size at the 5% level is 2.8*se; where se is the standard error of the effect size. The MDE to have 80% power to detect an effect size at the 0.5% level is 3.65*se.

Multiple Testing Problem

Multiple testing of for instance many outcome variables in an experiment, many treatments, or many different specifications to test the same hypothesis increases the likelihood of by chance finding a ”significant finding” (i.e. a false positive).

E.g. if 20 different true null hypotheses are tested one will on average be significant by chance with a $p<0.05$ significance threshold.

Bonferroni correction; divides the significance threshold (e.g. 0.05) by the number of tests (e.g. 20 tests leads to 0.05/20=0.0025 significance threshold). Conservative if the tests are correlated.

For testing several different hypotheses explicitly incorporating the prior is an alternative (but difficult to determine the prior).

P-hacking and ”Forking”

Researcher degrees of freedom; e.g. choice of statistical test, choice of covariates (and functional form), choice of outcome variable, choice in how to handle outliers, choice in sub-group analyses, choice in when to stop data collection.

P-hacking: a more conscious process of testing different variants and report the most significant (or testing variants until a significant one is found).

Forking: a more unconscious process of making choices that favors the tested hypothesis.

Is p-hacking/forking likely to be most prevalent in experimental studies or studies based on observational data?

Preventing P-hacking/Forking: Preregistration

Preregistration: Posting an analysis plan (preanalysis plan) prior to collecting the data (detailing the hypotheses and tests to be conducted in advance; including all the choices made during the analysis on statistical test, covariates, outliers, etc).

A distinction can be made between primary hypotheses, secondary hypotheses and exploratory analyses in the preanalysis plan.

Can be posted at for instance OSF (Open Science framework: https://osf.io/)

Some journals publish Registered Reports (e.g. Nature Human Behaviour); where a pre-analysis plan is submitted to the journal and the publication decision is based on the pre-analysis plan prior to starting the data collection.