Experimetrics: A Survey

Peter G Moffatt
24 November 2020

ABSTRACT

This monograph aims to survey a range of econometric techniques that are currently being used by experimental economists. It is likely to be of interest both to experimental economists who are keen to expand their skill sets, and also the the wider econometrics community who may be interested to learn the sort of econometric techniques that are currently being used by Experimentalists. Techniques covered range from the simple to the fairly advanced. The monograph starts with an overview of treatment testing. A range of treatment tests will be illustrated using the example of a dictator-game giving experiment in which there is a communication treatment. Standard parametric and non-parametric treatment tests, tests comparing entire distributions, and bootstrap tests will all be covered. It will then be demonstrated that treatment tests can be performed in a regression framework, and the important concept of clustering will be explained. The multilevel modelling framework will also be covered, as a means of dealing with more than one level of clustering. Power analyisis will be covered from both theoretical and practical perspectives, as a means of determining the sample size required to attain a given power, and also as a means of comuting ex-post power for a reported test. We then progress to a discussion of different data types arising in Experimental Economics (binary, ordinal, interval, etc.), and how to deal with them. We then consider the estimation of fully structural models, with particular attention paid to the estimation of social preference parameters from dictator game data, and risky choice models with between-subject heterogeneity in risk aversion. The method maximum simulated likelihood (MSL) is promoted as the most suitable method for estimating models with continuous heterogeneity. We then consider finite mixture models as a way of capturing discrete heterogeneity; that is, when the population of subjects divides into a small number of distinct types. The application used as an example will be the level-k model, in which subject types are defined by their levels of reasoning. We then consider other models of behaviour in games, including the Quantal Response Equilibrium (QRE) Model. The final area covered is models of learning in games.

CONTENTS

1. Introduction

Experimetrics[1] comprises the body of econometric techniques that are customised to experimental applications. This monograph is aimed principally at two types of reader. The first is the experimental economist who is interested in expanding their skill set in econometric techniques. In particular, experimental researchers who rely heavily on straightforward treatment testing techniqes will hopefully be persuaded that more sophisticated econometric techniques are more suitable in many settings. The second type of target reader is the researcher from the wider econometrics community who may be interested to discover the sort of econometric techniques that are currently being used by Experimentalists. Of particular interest to this type of reader may be the techniques that have not traditionally been part of the standard Econometrics toolkit, for example, power analysis and optimal experimental design. Possibly the best possible end result from this Monograph would be the cross-fertilisation of ideas from Applied of Theoretical Econometrics resulting in new and improved techniques in Experimetrics.

The most widely used approach in the experimental economics literature is the treatment testing approach. The treatment test often simply amounts to a comparison of some behavioural outcome between two samples: a control sample and a treatment sample, drawn randomly from the population of experimental subjects. Justification of this approach is usually provided in terms of subjects being assigned randomly to treatments, and all influences on behaviour other than the treatment of interest being held fixed by virtue of the experimental environment. These arguments are often used to explain away many of the types of problem that econometricians have traditionally been interested in, such as sample selection bias, measurement error, and endogeneity.

However, there are a number of compelling reasons why the level of econometrics required in the analysis of experimental data goes beyond simple treatment testing. Firstly, in planning a treatment test, issues of experimental design are important. The most basic feature of the design is the sample size, and the use of power analysis is becoming routine in the process of setting this design feature. Power analysis has underpinnings in statistical theory, and conducting a thorough power analysis is facilitated by an understanding of this underlying theory. It must be said that this interest in the use of power analysis is relatively new in econometrics and has taken hold mainly as a consequence of the rise of Experimental Economics.

Second, there are many different ways of conducting the treatment test itself. A choice must be made over parametric and non-parametric tests, between-subject and within-subject tests, and so on. Which approach is best suited in a particular situation is often far from obvious. Once again power analysis is important. Testing approaches vary widely in terms of power. However, a theme of this monograph is that there are other considerations relating to human behaviour which sometimes distort theoretical prescriptions, and the test with the highest power is not always the most suitable.

[1] The word "Experimetrics" was (to the best of my knowledge) coined by Camerer (2003, p.42). The first article containing the word in the title was Bardsley and Moffatt (2007), and a textbook bearing the title *Experimetrics* was produced by Moffatt (2015).

Third, the structure of experimental data is frequently such that straightforward application of treatment tests is invalid. Data is often clustered, sometimes at more than one level. It is very common for each subject to engage in a sequence of experimental tasks, and hence there is clustering (or dependence) at the subject-level. In some types of experiment, subjects interact in groups, and it is natural to expect clustering at the group level. Finally, an experiment may be divided into sessions taking place in different places or at different times, and we may expect clustering at the session level. It is well-known (Moulton, 1986) that any combination of these different levels of clustering invalidates any test that assumes independence between observations. One way of dealing with these problems is to conduct the treatment tests as tests of significance in regressions with clustered standard errors, but a superior approach is to use a multilevel model that fully incorporates all levels of clustering.

Fourth, most treatment tests are performed on the assumption that if agents respond to a treatment, they all respond to it in the same way. That is, there is an implicit assumption of homogeneity. It may be that a proportion of subjects respond to the treatment in the expected way, but the remainder are not affected by the treatment. In this case, a test that assumes homogeneity will underestimate the treatment effect for those who respond to it. A more extreme possibility is that half of the population respond positively to the treatment while the other half respond negatively, and in this case the homogeneous treatment test may completely fail to detect any effect.

The importance of subject heterogeneity extends far beyond treatment effects. In many experiments, the focus is on "home-grown" characteristics of experimental subjects, such as risk aversion, ambiguity aversion, discount rate, inequity aversion, aversion to lying, and depth of reasoning. All of these characteristics are likely to exhibit wide variation over the population of subjects. Any model of any aspect of behaviour in which such characteristics are relevant must allow for this variation. Often, the most natural way to allow for such variation is to construct a fully structural model in which the distributional parameters of the varying characteristics are estimated along with the treatment effects of interest.

A good example of this is the modelling of social preference data. In some settings, individuals face repeated tasks in each of which they decide how to divide an endowment between theirself and and another player, with the endowment and price of transferring varying between tasks. This sort of data set provides an opportunity to estimate social preference parameters such as aversion to inequity and preference for efficiency – parameters which are almost certain to exhibit between-subject heterogeneity.

Another example is the modelling of risky choice data. Here, data on repeated choices between lottery pairs may be used to estimate preference parameters such as risk aversion and probability weighting. Again, these are parameters for which between-subject heterogeneity is expected. This sort of heterogeneity is allowed for in estimation using the method of maximum simulated likelihood.

Some sorts of heterogeneity may be referred to as *discrete heterogeneity*, meaning that, instead being characterised by continuously varying preference parameters, it takes the form of the population subjects dividing into discrete "subject-types", with discretely different models of behaviour. One example is behaviour in public goods games, in which the population might be assumed to divide between free-riders (who follow the Nash prediction by contrinuting zero), reciprocators (who are willing to contribute only if they see others contributing), selfish contributors (who contribute in anticipation of reciprocity by others), and altruists (who contribute regardless). Note that these four types are each defined by a

different sub-model, and the econometric objective is to use the experimental data to estimate the parameters of these four models, along with a set of mixing proportions, which reveal the proportions of the population who are of each type. The econometric framework used for this purpose is the finite mixture model. A variety of methods are available for the estimation of finite mixture models, including machine learning techniques.

Another application of the finite mixture model which is used for illustrative purposes in this monograph is to depth of reasoning models. Here, the objective is to use data from behaviour in one-shot interactive games in order to estimate the proportion of the population who act at each level of reasoning.

Also of interest is behaviour in repeated games, where the initial questions to be addressed are how closely subject choices adhere to the Nash-equilibrium prediction, and whether sequences of choices appear random. More sophisticated analysis of repeated-game data reveals whether and how decision-makers *learn* to optimise their behaviour in repeated games. Are they attracted to strategies that have proved beneficial in previous tasks (reinforcement learning), or do they go further and form beliefs about the other players' actions and then make optimal decisions based on these beliefs (belief learning)?

The purpose of this monograph is to survey these techniques. However, it aims to be more than just a survey of the literature because it also aims to explain the techniques and evaluate them with illustrations, sometimes using data sets from previously published studies.

Section 2 provides an overview of treatment testing. A range of treatment tests are illustrated using the example of a dictator-game giving experiment in which there is a communication treatment. Standard parametric and non-parametric treatment tests, and also tests comparing entire distributions, are surveyed briefly. Then bootstrap tests are proposed as a way of avoiding the disadvantages of both parametric and non-parametric tests. Then it is explained how a treatment tests can be performed in a regression framework, as the test of significance of the effect of a dummy variable representing the treatment, one considerable advantage of this approach being that it becomes possible to correct the test for clustering of the data that inevitably arises at the subject-level and/or the session level. The final part of the section covers multi-level modelling, in which subject-specific and session-specific random effects are both incorporated in estimation.

Section 3 covers power analysis, an area of growing importance in experimental economics. The question of central interest is usually: what sample is required in a treatment test to attain a benchmark level of power (e.g. 0.8)? Tailor-made routines are outlined, but it is made clear that these methods are only applicable to a fairly limited range of treatment testing problems. The Monte Carlo method is proposed as a useful method for performing power calculations in situations in which ready-made routines are not available. The method is be applied to investigate questions including: how much more powerful is a within-subject test with n subjects than a between-subject test with 2n subjects? Then the method is applied to consider the problem of how to choose both n and T simultaneously in order to attain a required power in a panel data context.

Section 4 covers the different types of data arising in Experimental Economics, and the reasons why they arise in the form they do. Decision-making under risk is the chosen context in which these are considered. This choice of context is convenient because different data types arise depending on the elicitation method used: binary data arises when the subject chooses between lotteries; ordinal data arises when the subject makes a choice and also reports

strength of preference; interval data arises when the subject is faced with a sequence of choice problems and decides on their switch-point, that is, where in the sequence they switch from the safe to the risky choice or vice versa; exact data arises when the subject has been asked to report a certainty equivalent of a single lottery.

The remainder of the monograph is concerned mainly with the estimation of fully structural models. Section 5 considers methods for the structural estimation of social preference parameters. Section 6 considers ways of dealing with continuous heterogeneity. The method of maximum simulated likelihood (MSL) is be used for all examples, and this method is explained in reasonable detail. Models with one dimension of heterogeneity, such as the Fechner and RP models of risky choice, in which only risk aversion varies between subjects, are be considered first. We then move on to settings in which there is more than one dimension of heterogeneity, for example: models in which subjects vary in both risk aversion and probability weighting; models in which subjects vary in risk aversion and time preference; models in which subjects vary in risk aversion and inequity aversion.

Section 7 considers finite mixture models as a way of capturing discrete heterogeneity; that is, when the population of subjects divides into a small number of distinct types. The application used as an example will be the level-k model, in which subjects are defined by their levels of reasoning, with the level of reasoning typically ranging from 0 up to 3 or 4. Section 8 considers other models of behaviour in games, including the Quantal Response Equilibrium (QRE) Model. Section 9 considers models of Learning. Section 10 concludes.

2. Treatment Testing

No survey of Experimetrics would be complete without a section on treatment testing. This is because a surprisingly large part of Experimental Economics relies exclusively on this approach in the analysis of data and the testing of hypotheses. The essence of the approach is very simple, and normally amounts to the comparison of two data samples. However, within the area of treatment testing, many different methods are available to choose from: between-subject versus within-subject; parametric versus non-parametric; and so on. Hence a key question to be addressed is which method should be used in any given situation.

Another set of questions intimately related to treatment testing are those of power analysis. Power analysis is the set of techniques used to compute the statistical power of a treatment test in a given situation, and perhaps more usefully in the context of Experimental Economics, to find the sample size that is required to attain a given required power. One purpose of this section is to prepare the ground for Section 3 which is on the topic of power analysis.

2.1. Key Concepts and definitions

Here we provide a brief review of the central concepts underlying treatment testing, and in doing so we introduce the terminology that is used in this section and again in Section 3 when power analysis is discussed. For further detail on these concepts see Siegel and Castellan (1988).

We assume that the experiment consists of two treatments, 1 and 2. One of the treatments (treatment 1 say) is sometimes referred to as “control”. We are interested in the value of a scalar valued parameter θ. Let us assume that this parameter has (population) value θ_1 under

treatment 1 and θ_2 under treatment 2. The difference between the parameter values under the two treatments is $\delta = \theta_2 - \theta_1$, and δ is known as the "true effect size".

A treatment test always has a *null hypothesis,* labelled H_0, and an *alternative hypothesis*, labelled H_1. The null hypothesis is typically the hypothesis that there is no effect, that is, the true effect size, δ defined above, equals zero. The alternative hypothesis is that there is an effect. If the alternative hypothesis specifies the direction of the effect (i.e. $\delta > 0$ or $\delta < 0$), it is a *one-sided alternative* and we conduct a *one-tailed test*. If the alternative does not specify the direction of the effect (i.e. $\delta \neq 0$) it is a *two-sided alternative* and we conduct a *two-tailed test*. One-sided alternatives are usually proposed when the researcher has a prior belief about the direction of the effect, the prior belief perhaps coming from economic theory, or some behavioural theory.

In some situations, for example when performing power calculations, it is necessary to fix the effect size under the alternative at a particular value. Hence the test becomes a test of H_0: $\delta = 0$ against H_1: $\delta = \delta_1$ where δ_1 is some pre-determined non-zero value, perhaps obtained from the results of a previous study.

The first stage of the application of the hypothesis test is to compute the *test statistic* which is a function of the *n* data values in the sample. *n* is the sample size. The second stage is to 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). The tails of this distribution form the rejection region of the test, and if the test statistic falls in this region, the null hypothesis is rejected in favour of the alternative. If the test statistic falls elsewhere, the null hypothesis is not rejected, and it may be concluded that the test result is consistent with the null hypothesis.

The rejection region is determined by whether the test is two-tailed or one-tailed, and by the chosen *size* of the test. The size, usually denoted as α, is the probability of rejecting the null hypothesis when it is true, and this is normally set to 0.05. The point at which the rejection region starts is referred to as the critical value of the test.

The p-value of the test is the probability of obtaining a test statistic that is at least as extreme as the one obtained. One reason why the p-value is useful because it allows a conclusion to be drawn without comparing a test statistic to a critical value (i.e. it avoids the need to consult statistical tables). A more important reason why the p-value is useful is because it is a measure of the strength of evidence against the null, and in favour of the alternative (i.e. the strengh of evidence of an effect). The words used to represent strength of evidence are a matter of individual taste. According to popular convention: if $p<0.10$, there is mild evidence of an effect; if $p<0.05$, there is evidence; if $p<0.01$, there is strong evidence; if $p<0.001$, there is overwhelming evidence.

As mentioned above, a prior belief about the direction of an effect leads to a one-tailed test. For a one-tailed test (assuming the test statistic has the expected sign) the p-value is half of the p-value for the corresponding two-tailed test. Hence one-tailed tests are more likely to find evidence of an effect. In this sense prior beliefs are very useful because they can be used to boost the chances of obtaining a conclusive result.

Rejecting the null hypothesis when it is true is known as a type 1 error. As noted above, the probability of a type 1 error is denoted as α, and is usually set to 0.05. The other type of error

is a type 2 error: failing to reject the null hypothesis when it is false. The probability of a type 2 error is denoted as β.

The *power* of a test is the probability of rejecting the null hypothesis when it is false. The power is denoted as π. Note that π=1-β. The power of a test is determined by a number of factors, including the true effect size, the sample size (n), and whether the test is one-tailed or two-tailed. It also depends on the chosen value of α. The higher α is, the higher the probability of type 1 error, which has the benefit of higher power. *Power analysis* is the name given to the set of techniques used to compute the power of a given test, and to find the sample size required to meet a given power requirement.

Because the power of a test depends on the true effect size, and the true effect size is typically unknown, power is often depicted graphically using a power function. This is power plotted against true effect size. It is also informative to plot power against sample size, for a given true effect size. Examples of such plots will be presented in Section 3.2.

2.2 Choosing the values of α and β

The close relationship between size and power clearly implies that the choice of size (α) is an important decision. This choice depends to a large extent on the type of hypothesis under test. For a well-known example, first consider a situation in which the null hypothesis is that a crime suspect is not-guilty, and the alternative is that the suspect is guilty. In this situation, a type 1 error is finding an innocent person guilty, while a type 2 error is letting a guilty person go free. Many people view the first error as more serious than the second. Hence we should choose a very low value of α in this situation. How low? This is another question, although it must be noted that α cannot be lowered all the way to zero, since this would mean that every suspect must be declared not-guilty.

For a second well-known example, that is particularly topical in 2020, consider a situation in which the null hypothesis is that a patient is healthy, and the alternative is that they are suffering from a contagious disease. In this situation, declaring a diseased patient healthy (type 2 error) might be considered much more serious than telling a healthy patient that they have the disease (type 1 error). Hence, in this situation we could allow a higher value of α, since this would give rise to a higher probability of detecting infected patients (i.e. higher power).

In Experimental Economics, we are perhaps fortunate that the errors that might be made in the interpretation of results from hypothesis tests rarely have consequences that are as serious as those arising in the above examples. Hence it seems reasonable to follow convention and set the value of α to 0.05 as a standard. Closely related to the choice of α are a number of important recommendations - relating to the "scientific quality debate" – that should be followed in the conduct of tests. These recommendations are discussed in Section 3.3 below.

Although there are no formal standards for power, many researchers assess the power of their tests using π=0.80 as a standard for adequacy. The corresponding value of β is 0.2. These conventions imply a four-to-one ratio between the probability of type II error and the probability of type I error. So, returning to the example of the suspect in court, the number of guilty suspects set free should be approximately four times the number of innocent suspects imprisoned.

2.3 Treatment tests illustrated

In this Section, we will demonstrate a number of standard treatment testing techniques in a context that is very popular in Experimental Economics: dictator game giving. The dictator game was first introduced by Forsythe et al. (1994) with the objective of investigating how much individuals are willing to give to others, and identifying the determinants of giving behaviour. The game has two players, A (dictator) and B (recipient). Player A has the task of dividing some fixed endowment ($100 say) between themselves and Player B. Player A's decision takes the form: I get $90; you get $10. The role of Player B is entirely passive. Observed behaviour in this game allows a very straightforward test of the *homo economicus* model of individual behaviour: if individuals were only concerned with their own economic well-being, Player A would allocate the entire amount to themselves and transfer zero to Player B. In game theoretic terms, zero giving is the Nash Equilibrium prediction in this game. There is a vast amount of experimental evidence to reject this model: on average dictators give about 20% of the endowment to the receiver (Camerer, 2003).

We are clearly interested in why dictators depart from the theoretical prediction of zero giving. We are also interested in ways in which giving behaviour may be manipulated by altering the setting. The vast literature on the Dictator game has not only found that some individuals give to others, but also that giving behaviour is highly sensitive to certain design features. Engel (2011) provided a thorough meta analysis of dictator games, pooling the results of more than 100 studies. Factors he found to have a positive effect on giving include "endowment earned by recipient" and "deserving recipient". Factors having a negative effect include "endowment earned by dictator" and "repeated game".

For the purpose of this monograph, we are not so much interested in the effects of design features on giving, as simply in the statistical methods used in individual studies in order to test these effects. Some specific examples are useful at this point. Cherry et al. (2002) focus on the effect of whether the endowment is earned by the dictator, using Fisher's exact test to analyse the effect on the decision of whether or not to give, and using the Mann-Whitney test for the effect on the amount given. Eckel and Grossman (1998) focus on the effect of the gender of the dictator, and use the independent samples t-test to compare the means of giving by gender, and the Kolmogorov-Smirnov and Epps-Singleton tests to compare the entire distribution of giving by gender. Bardsley (2008) investigates the effect of a "taking treatment" on the propensity to give, using a bootstrapped difference-in-proportions test.

Even on the basis of this small number of examples, it is clear that a fairly wide range of tests are available and have been used for the purpose of treatment testing. Here, the objective will be to demonstrate and evaluate each of these tests in a chosen context, with a fixed test data set. The example chosen for illustration of treatment testing is communication. The question of interest is: does the behaviour of the dictator change if the responder is permitted to communicate with them, by sending a message before the giving decision is made? Although such a message must be classified as "cheap talk" and does not alter the theoretical prediction of zero giving, there is clearly a potential for communication to influence behaviour.

There are broadly two ways of testing whether communication has an impact on giving. An independent-samples test assigns some dictators to a control group (no communication) and other dictators to a treatment group (with communication), and then investigates the

difference in behaviour between the two groups.[2] A within-subject test carries out both procedures on each dictator. That is, each dictator's behaviour is observed both without and with the treatment. Again, behaviour is compared between treatments.

The application is particularly useful for the demonstration of treatment tests because a range of different types of test are required, and can therefore be demonstrated naturally.

2.4 Dictator game data

Let us assume that a dictator game experiment has been carried out on a sample of subjects. The dictator's endowment is $100, and the decision variable is the number of dollars transferred from the dictator to the receiver. Let us assume that the giving data is as presented in Table 2.1.[3] Since the purpose is simply to illustrate the use of key tests, this data set is simulated, in such a way as to exhibit the key stylised facts of dictator game experiments.

	y1	y2
1.	0	0
2.	20	20
3.	11	12
4.	31	28
5.	0	0
6.	22	26
7.	27	43
8.	3	10
9.	10	25
10.	17	20
11.	6	16
12.	36	36
13.	1	3
14.	50	50
15.	0	0
16.	22	19
17.	1	0
18.	5	17
19.	0	0
20.	18	22
21.	0	0
22.	0	50
23.	0	0
24.	28	32
25.	21	26
26.	26	50
27.	0	0
28.	30	28
29.	30	50
30.	0	0

Table 2.1: Simulated Dictator game giving for 30 dictators. y1 is giving with no communication; y2 is giving with communication.

The simulated data set presented in Table 2.1 is from a within-subject experiment. That is, there are 30 dictators, and each dictator has performed the task twice, with a different recipient on each occasion. On the first occasion (y1), the dictator and recipient have no opportunity to communicate; on the second occasion (y2), they have such an opportunity.

[2] Note that only half of the subjects in a Dictator Game task are dictators. The other half are recipients. Although recipients have no active role in the experiment, their participation is essential for the dictators' task to have meaning.

[3] The simulated data is presented here for the benefit of readers who wish to reproduce the results reported.

Although this data is within-subject data, it will be used to illustrate both within-subject tests and independent-sample tests. For the independent-sample tests, the pairing of observations will simply be disregarded, and the two colums of Table 2.1 will be treated as two independent samples. Disregarding the pairing is clearly inadvisable since valuable information is being ignored. However, using the same data set for both within and between tests allows a sharp comparison between the outcomes of the two types of test, and in particular allows an assessment of the advantages from using paired data over independent samples.

For the independent-sample tests, the data will therefore be treated *as if* there are 30 dictators in each group, with the first column showing the giving behaviour of the 30 dictators in the control group (no communication), and the second column showing that of the treatment group (communication).

Table 2.2 presents descriptive statistics for the data shown in Table 2.1. We find that the sample means of the two columns (Control and treatment) are \$13.83 and \$19.43 respectively. The key question to be asked is whether this difference is statistically significant.

It is useful to start by plotting the two distributions in histograms. This is done in Figure 2.1. This is useful because it gives clues to the precise nature of the treatment effect. We see that although the treatment appears to have little effect on the number of zero-givers, there appears to be a shift to the right among positive givers, with a clear increase in the number of "equal splitters".

	n	***mean***	***sd***
No communication	30	13.83	14.04
Communication	30	19.43	17.45
Pooled	60	16.63	15.95
correlation	0.7830		

Table 2.2: Descriptive statistics for simulated sample presented in Table 1

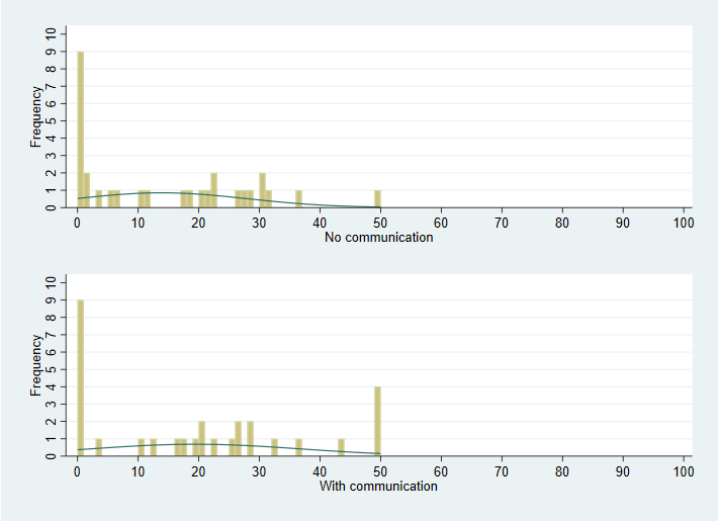

Fig. 2.1: Frequency histograms of (simulated) dictator game giving without (n_1=30) and with (n_2=30) communication. Normal densities superimposed.

Histograms of the type shown in Figure 2.1 are useful for comparing two independent samples. Given that the data shown in Table 2.1 is actually paired data, it is perhaps more informative to obtain a scatter plot of the second column against the first. This is done in

Figure 2.2. The observation that most of the points lie above the 45-degree line is consistent with the hypothesis that communication increases giving.

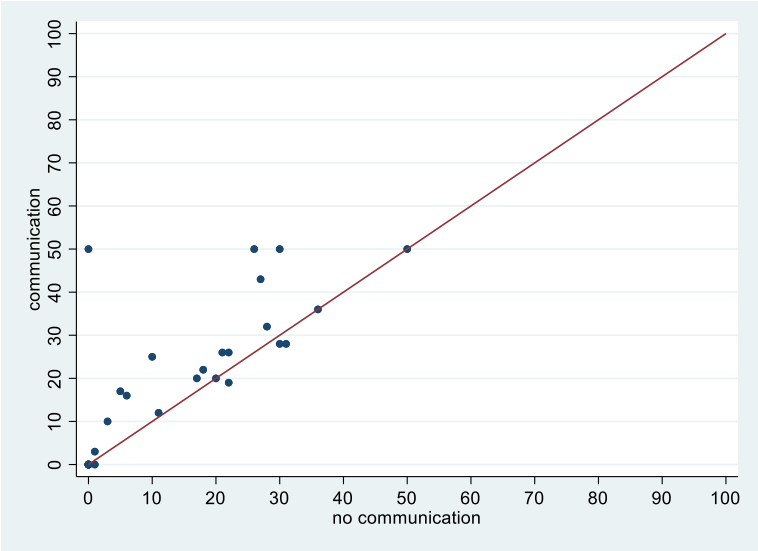

Figure 2.2: Giving in communication treatment against giving in no-communication treatment. Paired data (n=30). 45-degree line superimposed.

Test	***Test statistic***	***p-value***	***Conclusion***
Normality tests:			
Skewness		0.0525	Mild evidence of skewness
Kurtosis		0.2405	No evidence of non-normal kurtosis
Joint (sktest)	$\chi^2(2)=5.05$	0.0802	Mild evidence of non-normality
Shapiro-Wilk test	Z=2.423	0.0077	Strong evidence of non-normality
Independent-sample tests:			
Indep samples t-test	t(58)=-1.3956	0.0881	Mild evidence of difference in means
Sd test	F(29,29)=0.6480	0.2486	No evidence of difference in variances
t-test (unequal variances)	t(.) = -1.3695	0.0882	Mild evidence of difference in means
Permutation t-test	t=-1.3956	0.0916	Mild evidence of difference in means
Mann-Whitney test	z = -0.907	0.1821	No evidence of difference in medians
Bootstrap t (B=9,999)		0.0895	Mild evidence of difference in means
Kolmogorov-Smirnoff	0.1167	0.435	No evidence of different distributions
Epps-Singleton		0.5643	No evidence of different distributions
Chi-squared test (+ vs 0)	$\chi^2(1)=0.30$	0.584	No difference in propensity to give
Fisher's exact test	N/A	0.392	No difference in propensity to give
Within-subject tests:			
Paired t-test	t (29) = -2.8248	0.0042	Strong evidence of difference in means
Wilcoxon test	z = -2.929	0.0034	Strong evidence of difference in medians
Sign test		0.0096	Strong evidence of difference
McNemar test	$\chi^2(1)=0$	1	No difference in propensity to give positive amount

Table 2.3: Results from tests applied to simulated dictator game data presented in Table 2.1. For the independent-sample tests, the pairing of observations is disregarded, and the two columns of data are treated as two independent samples of size 30.

Table 2.3 presents the results from a range of tests applied to the simulated data shown in Table 2.1. Obviously, some of these tests are more appropriate than others, and the choice of which test to use depends on the experimental design, the nature of the data, and the research question. The purpose of Table 2.3 is to present the range of tests that are typically applied in the Experimental Economics literature, when the focus of analysis is a single treatment effect. Perhaps the most striking feature of Table 2.3 is that most of the within-subject tests all result in strong evidence of a treatment effect, while the independent-sample tests result in at most mild evidence.

In the remainder of Section 2, we consider each of these tests in more detail, and in Section 3 we go further by investigating their power properties. There, we find that within-subject tests tend to have considerably higher power than independent-sample tests, and this is consistent with the differences seen in Table 2.3.

2.5 Tests of Normality

An important issue to address is whether the data is normally distributed. Referring back to Figure 2.1, we see that neither of the two distributions appears close to the superimposed normal densities. This is partly a consequence of the significant accumulations of observations zero (the Nash prediction) and at 50 (equal splitting of the endowment). Although the skewness-kurtosis test (see Table 2.3) finds only mild evidence of non-normality,[4] the Shapiro-Wilk test finds strong evidence (p=0.0077) to reject normality.

The question of normality is important because the independent samples t-test, considered in the next section, is based on the assumption of normality of the two distributions whose means are being compared. The test may not be valid if the distributions are non-normal. Of course, it is conventional to appeal to the Central Limit Theorem (CLT) when the samples being compared are sufficiently large. According to the CLT, the (standardised) mean of a sufficiently large sample follows a standard normal distribution even when the sample is drawn from a sample that is not normal (see e.g. Berenson et al., 1988). Here "sufficiently large" is often considered to be 30 and above. This requirement is (marginally) met in the present case, so it may be that we are justified in relying on the t-test result.

2.6 Parametric (independent-ssample) Treatment Tests

As explained in Section 2.4, for the purpose of demonstrating independent-sample tests, the pairing of the data in Table 2.1 will be disregarded, and the data will be treated *as if* there are 30 dictators in each of two groups, with the first column showing the giving behaviour of the 30 dictators in the control group (no communication), and the second column showing that of the treatment group (communication).

We commence with the independent samples t-test of the null hypothesis that the two means are equal. We will refer to the two population means as μ_2 and μ_1, and hence the null hypothesis under test is $H_0 : \mu_2 = \mu_1$, and the alternative is $H_0 : \mu_2 > \mu_1$. Note importantly

[4] It is often found that tests of normality appear deficient in power.

that the alternative hypothesis is one-sided since our behavioural model predicts that communication has a positive effect on giving. The test statistic is given by:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.1}$$

where s_p is the pooled standard deviation. The test statistic (2.1) follows a $t(n_1 + n_2 - 2)$ distribution under H_0, and hence H_0 is rejected in favour of H_1 if $t > t_{n_1+n_2-2,0.05}$.

When applied to the present data set, this test results in a t-statistic of +1.36 and a one-tailed p-value of 0.088, leading us to conclude that there is only mild evidence that communication increases giving.

The version of the independent samples t-test performed above is based on the assumption that the two distributions being compared have equal variances. This can be checked using the variance-ratio test:

$$F = \frac{s_2^s}{s_1^2} \tag{2.2}$$

The test statistic F in (2.2) has a F_{n_2-1,n_1-1} distribution under the null hypothesis of equal (population) variances. Here, we find that F=1.54, with (2-tailed) p-value 0.25. This indicates that we have no reason to believe the variances to be unequal, and that we are justified in using the equal-variances version of the t-test.

2.7 Non-parametric (independent-sample) Treatment Tests

In the interests of conservativeness, we shall assume that we have reason to believe that the t-test considered in Section 2.6 may be unreliable. Given this assumption, the natural alternative is a non-parametric test. Perhaps the most popular non-parametric treatment test among Experimental Economists is the Mann-Whitney test. It is classified as non-parametric because it does not rely on any strong distributional assumptions (such as normality of the data). It must be recognised however, that the test does depend on the assumption of equal variances between the two populations (which has already been checked for the present sample).

A useful way of viewing the Mann-Whitney test is as a comparison of the *medians* of two samples, as distinct from the independent samples t-test which is based on the comparison of two *means*.

To implement the Mann-Whitney test, all of the observations from both samples are ranked by their value, with the highest rank being assigned to the largest value, and with ranks averaged in the event of a tie. Then the sum of ranks are found for each sample, and compared. The test is based on this comparison. See Siegel and Castellan (1988) for further detail.

For the test sample, a z-statistic of 0.907 is computed from the rank sums, and the (one-tailed) p-value is 0.1821. It appears that there is no evidence of a treatment effect on the basis of this test. The fact that the evidence of a treatment effect is weaker with the Mann-Whitney

test than with the t-test (reported above as p=0.0882) is an expected result: evidence of an effect tends to be weaker, the less is assumed about the process generating the data.

2.8 The Bootstrap

In Section 2.7 the Mann-Whitney test was introduced as a non-parametric analogue to the two-sample t-test used in Section 2.6, indicating that the former may be preferred in situations in which one doubts the assumption of normality of the data. However, one drawback of non-parametric tests of this type is that they are based solely on the *ordinality* of the data, and hence they completely disregard the (possibly) rich *cardinal* information in the data.

The "bootstrap" technique (Efron and Tibshirani, 1993) provides a means of conducting a parametric test such as the two-sample t-test (which definitely respects cardinality), without making any assumptions about the distribution of the data.[5] To apply the bootstrap procedure to the t-test requires the following five steps:

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

According to Mackinnon (2002) the number of bootstrap samples, *B*, should be chosen so that $\alpha(B+1)$ is a whole number, where α is the chosen test size. Since α is usually set to 0.01, 0.05, or 0.10, this requirement essentially means that *B* should be either 99 or 999 or 9999, etc. With B=9,999 bootstrap samples of the t-test on the example data set, the (one-tailed) p-value is 0.0895, which is slightly larger than the one obtained by applying the t-test directly (0.0881). Again, there is only mild evidence of a treatment effect.

An obvious drawback of the bootstrap approach is that it is a random procedure and therefore results in a different p-value each time the procedure is applied. The obvious remedy is to use a larger value of *B*. Another way of addressing this problem is to use a random number seed in the implementation of the bootstrap procedure. The chosen seed exactly determines the bootstrap samples generated, and is therefore a way of fixing the results of the bootstrapping procedure.

2.9 Tests comparing entire distributions

As suggested by Forsythe et al. (1994), tests comparing entire distributions are useful in situations in which economic theory does not predict the precise nature of the treatment effect. In the context of the current example, since (*homo economicus*) theory only predicts that the amount transferred by the dictator is zero, the same theory is not useful in predicting

[5] The range of applications of the bootstrap technique is very wide. It can be applied in the way described here to almost any test, and is also a standard method for obtaining measures of accuracy (standard error, bias, confidence interval, prediction error, etc.) to sample estimates.

the nature of any deviation from zero, and in particular it is not useful in predicting the nature of the impact of treatments such as communication. More precisely, theory does not predict which functional of the distribution may be expected to shift in response to the treatment. Is it (as usually assumed) the mean of the distribution that shifts? Or is it the median? Or is it the spread of the distribution? This problem is solved by applying tests that are based on a comparison of the entire distributions under the two treatments, rather than a comparison of a particular functional such as mean or variance.

One popular test that compares two entire distributions is the Kolmogorov-Smirnov (KS) test. In order to understand this test, it is useful to present the two cumulative distribution functions (cdf's) on the same graph. Such a graph is shown in Figure 2.3.[6] The observation that the with-cummunication cdf lies below and to the right of the no-communication cdf is consistent with giving being higher in the communication treatment. The KS test statistic is used to judge whether this difference is significant. With reference to Figure 2.3, the KS test statistic is seen to be something very simple: it is the maximum vertical distance between the two cdf's. This is in fact +0.1667. The corresponding p-value is found to be 0.435. Once again no treatment effect is found.

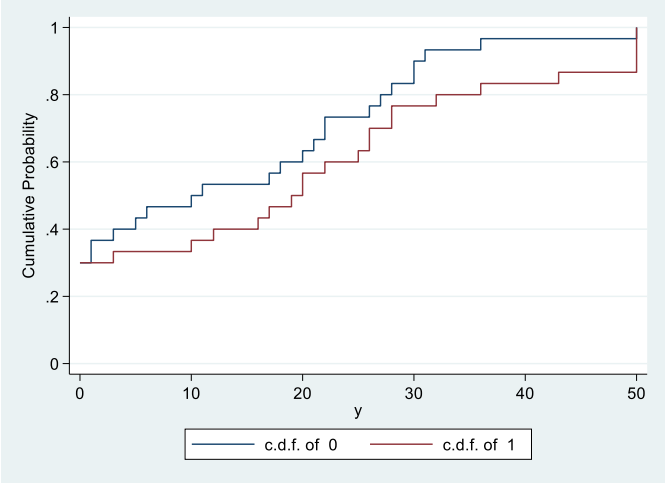

Figure 2.3: The cdf's of dictator's transfer under the no-communication treatment (higher line) and the communication treatment (lower line). The Kolmogorov-Smirnov test statistic is the maximum vertical distance between the two lines.

Another test for comparing entire distributions that has become popular among Experimental Economists is the Epps-Singleton test (Epps and Singleton, 1986). In fact, this test does not compare the two distributions directly, but instead compares the empirical characteristic functions.[7] This test is believed to perform similarly to the Kolmogorov-Smirnov test in terms of power, and has the added advantage of being applicable when the outcome has a discrete distribution (e.g. if the outcome is the number of questions answered correctly in a quiz). The test is implemented in STATA using the user-written command escftest (Georg, 2009). When applied to the present problem, the p-value is 0.56, indicating no evidence of a treatment effect.

2.10 Independent-sample tests with binary outcomes

[6] The graph shown in Figure 2.3 is obtained using the user-written STATA command cdfplot.

[7] For a sample of size *n*, $X_1 \ldots X_n$, the empirical characteristic function is defined as $\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(itX_j)$, where $i = \sqrt{-1}$.

In all tests discussed thus far, the outcome variable (giving) is continuously distributed. In this section we consider the situation in which the outcome is binary. This situation would arise if we were only interested in whether dictators give, and not how much they give. The decision variable is 1 if the dictator gives a positive amount, and zero otherwise. We then ask if this binary variable is associated with the binary treatment variable. The most popular test in this situation is the chi-squared test (see Siegel and Castellan, 1988). This test is based on the 2×2 cross-tabulation shown in Table 2.4. From the column percentages, we see that 63.33% of subjects give under no communication, while 70% give under communication. The chi-squared test essentially compares these two percentages. It does this by comparing the number in each cell with the number that would be expected if there were no treatment effect.

give positive	communication 0	1	Total
0	11 36.67	9 30.00	20 33.33
1	19 63.33	21 70.00	40 66.67
Total	30 100.00	30 100.00	60 100.00

Table 2.4. Cross-tabulation of positive giving indicator against treatment indicator, with column percentages.

For completeness, the chi-squared test statistic may be computed easily as (O stands for "observed" and E "expected"):

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(19-20)^2}{20} + \frac{(21-20)^2}{20} = \underline{0.30} \tag{2.3}$$

Under the null hypothesis of no treatment effect, the test statistic has a $\chi^2(1)$ distribution.[8] The test statistic of 0.30 computed in (2.3) does not fall in the upper tail of the $\chi^2(1)$ distribution. Hence there is no evidence of a treatment effect.

An alternative to the chi-squared test which is also popular in Experimental Economics is Fisher's Exact test. This test asks what is the probability of obtaining the combination of numbers in the tabulation, or a more extreme combination (i.e. a combination that favours the alternative hypothesis more), for the given row totals and column totals. This probability is the p-value for the test. For the above tabulation, the p-value is 0.392. Again we see no evidence of a treatment effect.

As the name suggests, Fisher's Exact test is an exact test, in contrast to the chi-squared test which relies on an approximation. Fisher's Exact test should be used in situations in which the

[8] The degrees of freedom for the chi-squared test is the number of entries in the cross-tabulation that are "free", given knowledge of the row sums and column sums. In a m×n cross-tabulation this is (m-1)×(n-1). Hence for a test involving a 2×2 tabulation, the degrees of freedom is 1.

chi-squared approximation is likely to fail, for example when the sample size is small, or when the numbers in some cells are very small.

2.11 Within-subject tests

In previous sections, it has been assumed that the two treatments in a treatment test have been administered to two samples separately. We referred to these tests as independent samples tests, but they are sometimes also called *between-subject* tests. *Within-subject* tests are used to test the effect of a treatment in the contrasting situation in which each subject is observed both with and without the treatment.

From a theoretical point of view, within-subject tests are preferred to between-subject tests, for the obvious reason that they have greater statistical power (see later Section 3 on power analysis). The obvious reason for this difference in power is that within-subject tests incorporate the additional information of the pairing of observations. However, there are various reasons why within-subject designs are not favoured by experimental economists. The issue of "order effects" is much discussed (see, for example, Harrison et al., 2005; Holt and Laury, 2005). An order effect is present if the result of the test depends on the order in which two treatments are administered. More generally, there are concerns that the experience of one treatment impacts on behaviour in the treatment that follows.

We will continue to use the same example data set used above to illustrate the independent-sample tests. However, we will now correctly recognise that there are only 30 dictators in total, and each dictator has been observed twice, without and with treatment. Although there are only 30 rows in the data set, it is sometimes useful to treat the data as a sample of 60 observations that are clustered[9] at the subject-level.

To test formally for a treatment effect, we may, as usual, choose between a parametric and a non-parametric test. The standard parametric test is the paired comparisons t-test. This 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. Applied to the test data, this results in a t-statistic of 2.82 and a p-value of 0.0042. This represents strong evidence that giving is higher in the communication treatment.

A non-parametric test appropriate in this situation is the Wilcoxon signed ranks test (see Siegel and Castellan, 1988). As with the parametric test, this 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. Then these ranks are summed separately for the positive differences and the negative differences. If there is no difference between the two treatments, these two rank sums should be roughly equal. The test is therefore based on a comparison of these two numbers. Applied to the test data, this results in a test statistic of z=2.929 and a p-value of 0.0034. This represents even stronger evidence of a treatment effect than that from the paired t-test.

The Wilcoxon signed ranks test is not completely distribution-free. It relies on the assumption that the distribution of paired differences is symmetric around the median. A test which avoids this assumption is the paired-sample sign test. This test simply 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. This test may be

[9] The concept of clustering is explained in more detail in Section 2.18 below.

viewed as a fully non-parametric test. The p-value from this test is 0.0096. The evidence from this test is still strong, but unsurprisingly weaker than the other two tests.

2.12 Within-subject tests with binary outcomes

In Section 2.11 we considered the problem of within-subject testing in a situation in which the outcome variable (giving) is continuously distributed. In this section we consider the situation in which the outcomes are binary.

If we are only interested in whether dictators give, and not how much they give, the outcome of interest is binary and the data can be presented in a 2×2 cross-tabulation, as in Table 2.5.

give positive no comm	give positive with comm: 0	1	Total
0	8	1	9
1	1	20	21
Total	9	21	30

Table 2.5: Cross tabulation for binary variables for giving positive amounts in two treatments. Within-subject data.

The question we are asking here is: is a particular subject more likely to give a positive amount with communication than without communication. To answer this question, we compare the two off-diagonal elements of the 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 in favour of our hypothesis. The test which formalises this comparison is in McNemar change test (see Siegel and Castellan, 1988). In this situation, the two elements in question are both 1, and hence the test will result in no evidence whatsoever in favour of the hypothesis. For this reason, it is appropriate to digress to more interesting applications of the McNemar test in other areas of experimental Economics. This is the purpose of the next sub-section.

The conclusion that there is no evidence of communication affecting the probability of giving a positive amount may at first sight appear at odds with the conclusions of Section 2.11, in which three tests unanimously found strong evidence that communication increases the amount given. This apparent contradiction suggests that a hurdle model might be appropriate. The hurdle model, originally due to Cragg (1971), separates the decision-making process into two stages: deciding whether or not to give; and deciding how much to give. The two stages are termed “hurdles”, and being a “zero-type” or a “selfish type” is characterised by “falling at the first hurdle”. This modelling framework is particularly useful in situations in which the two decisions are determined differently. Note that this is what precisely what we are seeing in the present example: the communication treatment has no effect on the probability of giving, but a large effect on the amount given.

The most interesting results from the estimation of hurdle models are those where an effect in the first hurdle is contradicted by an effect with the opposite sign in the second hurdle. One example of this contradiction is uncovered by Andreoni and Vesterlund (2001) who find that males are less likely than females to give a positive amount in dictator games, but that males who do give, tend to give more than females.

2.13 Within-subject tests with binary outcomes: other applications

The McNemar Change test (see Siegel and Castellan, 1988) is very popular in particular areas of Experimental Economics. One is the testing of Allais paradox (Allais, 1953), which is perhaps the most well-known contradiction of expected utility (EU) theory. The paradox is demonstrated by addressing a sequence of two (usually hypothetical) questions to a sample of subjects. The first question asks which they would prefer out of the lotteries A and A* below. The second question asks them to choose between B and B*.

Lottery A:	Certainty of $1 million
Lottery A*:	0.01 chance of nothing
	0.89 chance of $1 million
	0.10 chance of $5 million
Lottery B:	0.89 chance of nothing
	0.11 chance of $1 million
Lottery B*:	0.90 chance of nothing
	0.10 chance of $5 million

If a subject chooses A in the first question, and B in the second, we shall label their sequence of answers as "AB". There are clearly four different ways in which a subject can answer the two questions: AB, AB*, A*B, A*B*. Of these four possibilities, AB and A*B* are consistent with EU; AB* and A*B both indicate a violation of EU. In practice, a significant number of subjects do violate EU by choosing either AB* or A*B. However, what is of particular interest is the pattern, known as "Allais behaviour", of AB* violations being much more frequent than A*B violations.

In order to develop tests for the presence of Allais behaviour, we will use the notation n(.) to represent the number of subjects who answer with a particular sequence, for example, n(AB*) is the number of subjects who answer AB*.

The McNemar change test is conducted as follows. The null hypothesis is that AB* and A*B are equally likely. That is, we expect n(AB*) and n(A*B) to be approximately equal.[10] To test this null, we apply the chi-squared test to these two groups. The numbers in the other two groups are irrelevant. The test statistic is, after simplification:

$$\chi^2 = \frac{[n(AB^*) - n(A^*B)]^2}{n(AB^*) + n(A^*B)} \qquad (2.4)$$

The distribution of (2.4) is $\chi^2(1)$ under the null hypothesis of no Allais behaviour.[11]

Conlisk (1989) presented the two choice problems to 236 subjects. The numbers providing each response combination are given in Table 2.6. Applying (2.4) to the numbers in Table (2.6) yields a test statistic $\chi^2 = 63.6$, which, when compared to critical values of the $\chi^2(1)$ leads to the conclusion of overwhelming evidence of Allais behaviour.

[10] Note that we are assuming a simple error specification here, and a homogeneous population. Wilcox (2008, p.224-231) provides interesting examples in which populations satisfying EU can display Allais behaviour as a consequence of patterns of heterogeneity in which agents vary in "noisiness".

[11] If the two numbers being compared in the McNemar test are small, the approximation by the $\chi^2(1)$ distribution may become poor, since a continuous distribution is being used to approximate a discrete distribution. To deal with this problem, a "continuity correction" may be applied (see Yates, 1934).

	B	B*
A	18	103
A*	16	99

Table 2.6: Results from Conlisk's (1989) Allais experiment.

Conlisk (1989) suggested an alternative test statistic for detecting Allais behaviour, and this has come to be known as the Conlisk test. The statistic is given by:

$$Z = \frac{\sqrt{N-1}\left(S-\frac{1}{2}\right)}{\sqrt{\frac{1}{4V}-\left(S-\frac{1}{2}\right)^2}} \tag{2.5}$$

where N is the total number of subjects, V is the proportion of subjects who violate EU by giving AB* or A*B answers, that is:

$$V = \frac{n(AB^*)+n(A^*B)}{N} \tag{2.6}$$

And S is the proportion of violators who answer AB* rather than A*B, that is:

$$S = \frac{n(AB^*)}{n(AB^*)+n(A^*B)} \tag{2.7}$$

The test statistic (2.5) has a standard normal distribution under the null hypothesis of no Allais behaviour, with a value in the upper tail providing evidence that the proportion S given in (2.7) is significantly greater than one-half, that is evidence of Allais behaviour. Applied to the above data from Conlisk (1989), the test statistic is 9.32, which is certainly in the upper tail of the standard normal, and confirms the strong evidence of Allais behaviour in this sample.

Another setting in which the McNemar change test is popular is the testing of the preference reversal (PR) phenomenon (Grether and Plott, 1979). This is the phenomenon of subjects choosing the safer of two lotteries (the "p-bet") when asked to choose between them, but to contradict this choice by placing a higher valuation on the riskier lottery when asked to value them separately (that is, to provide a certainty equivalent for each).

As an example, consider the very first pair of lotteries considered by Tversky et al. (1990):

p-bet: 0.97 chance of $4; 0.03 chance of $0
$-bet: 0.31 chance of $16; 0.69 chance of $0

	Value p higher	Value $ higher
Choose p	43	106
Choose $	4	26

Table 2.7. Results from Tversky at al. (1990), study 1, set 1, triple 1.

There were 179 subjects, distributed as shown in Table 2.7. If a subject chooses p and values $ more highly, they are said to be making a "standard reversal". If they choose $ and value p

more highly, they are said to be making a "non-standard reversal". The PR phenomenon is detected if the number of subjects making standard reversals is significantly greater than the number making non-standard reversals.

With reference to Table 2.7, the number of standard reversals (106) is clearly much higher than the number of non-standard reversals (4). The two tests used above to detect Allais behaviour can be used here to formalise this comparison. The McNemar change test gives a $\chi^2(1)$ value of 94.58. The Conlisk test gives a Z-statistic of +14.07. Both represent overwhelming evidence of the PR phenomenon.

2.14 Treatment testing using regression models and multilevel models

Treatment tests can also be conducted in the context of a regression. In an independent-samples setting, if a regression is performed with the decision variable as the dependent variable, and with the treatment dummy as a single explanatory variable, and the regression includes an intercept, then the regression t-statistic associated with the treatment variable will be equivalent to the independent samples t-statistic discussed above.

The paired t-test can also be performed in the context of a linear regression. However, it is imperative to allow for the *dependence* between observations. A paired sample can be viewed as a collection of clusters of size 2. If a regression is performed with standard errors adjusted for this form of clustering, the result of the t-test will be very similar to that of the paired t-test discussed earlier.

In fact, one of the important advantages of the regression approach is that it allows this sort of adjustment for the dependence between observations. Dependence usually takes the form of clustering, either at the level of the individual subject (as with paired data), or at the level of the group of subjects that are interacting with each other, or at the level of the experimental session. It is well known that default OLS standard errors that ignore such clustering can greatly underestimate the true OLS standard errors, as emphasized by Moulton (1986).

Broadly, there are three ways of adjusting for dependence: ultra-conservative testing; clustering; and multilevel modelling. Ultra-conservative testing simply amounts to taking averages over the dependent observations at the highest level of clustering, and then performing treatment tests using this averaged data, which automatically satisfies independence. The glaring drawbacks of this approach are the neglect of the large amount of information in the individual-level data, and the greatly reduced sample size, inevitably leading to lower power. The term "clustering" is, in the treatment-testing context, usually taken to represent the process of obtaining "cluster-robust" standard errors for OLS estimates. These standard errors are obtained by assuming a block-diagonal covariance matrix for the regression error term. For a detailed discussion of the importance of clustering in Experimental Economics, see Frechette (2012). Multilevel modelling (Rabe-Hesketh and Skrondal, 2008) is a superior approach because it 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 multilevel model includes both the OLS model and the random effects model as special cases.

There are a number of other reasons why the regression framework might be the best approach to treatment testing. One is that it allows more than one treatment effect to be

tested at the same time. Another is that the regression context allows explanatory variables other than the treatment variable to be controlled for. For example, Ben-Ner et al. (2004) use dictator game data from 111 subjects, in which the focus of the analysis is the effect of the gender-combination of the two players. Using regression analysis, they find that female-to-female giving tends to be lowest, but they also control for personality traits such as religiosity, agreeableness, openness and cognitive ability.

Let us consider the multilevel model that is appropriate when there is dependence at the subject and group levels. Let y_{ijt} be the decision made by subject i in group j in task t. The model takes the following form:

$$\begin{gathered} y_{ijt} = \beta' x_{it} + \delta d_i + u_i + v_j + \varepsilon_{ijt} \\ i = 1, \cdots, n \quad j = 1, \cdots, J \quad t = 1, \cdots, T \\ V(u_i) = \sigma_u^2;\ V(v_j) = \sigma_v^2;\ V(\varepsilon_{ijt}) = \sigma_\varepsilon^2 \end{gathered} \tag{2.8}$$

In (2.8), d_i is the dummy variable indicating treatment,[12] and hence the parameter δ represents the treatment effect, x_{it} is a vector containing conditioning variables, u_i is the subject-specific random effect, v_j is the group-specific random effect, and ε_{ijt} is the observation-specific error term.

The model defined in (2.8) is a 3-level model. Note that if the highest level (group) is ignored, the model becomes a 2-level model, also known as a random effects model. If the highest level (group) and next highest level (subject) are both ignored, the model becomes a 1-level model, or just a linear regression model.

The multilevel modeling framework also allows random slopes. For example, if it is suspected that the treatment effect δ in (2.8) varies between subjects, it is possible to replace the subject-specific random effect u_i, by $u_{0i} + u_{1i}d_i$, where u_{1i} is the random component of the coefficient associated with d_i. Then $V(u_{1i})$ would be estimated and would indicate whether there is indeed heterogeneity in the treatment effect.

Gächter and Renner (2010) used a multilevel model of form (2.8) to test the effect of incentivising the elicitation of beliefs in public goods games, including psychological traits as conditioning variables. There were 204 subjects divided into 51 groups of 4, each facing 10 tasks. They found that when the elicitation of beliefs is incentivised (that is, when subjects are rewarded for accurately predicting the contributions of other group members), the subject tends to contribute more. Their explanation is that the incentivisation of beliefs has the effect of reducing the measurement error in beliefs, and hence enhancing the apparent effect of beliefs on contributions. Their estimates of the variance components are: $\sigma_u = 3.23$; $\sigma_v = 2.93$; σ_ε 4.50, suggesting that, although the equation error has the highest variance of the three error components, the between-subject and between-group variances also appear to be important, vindicating the choice of the multilevel model for their analysis.

3. Power Analysis

In recent years, Experimental Economists have become increasingly interested in the use of power analysis. Power analysis (Cohen, 2013) is used to find the power of a test that has been performed, power being defined as the probability of detecting an effect given that the effect

[12] Since the treatment variable d_i in (2.10) only has an i subscript, it represents a between-subject treatment. If the treatment were within-subject, there would also be a t subscript.

really exists. It is also used to find the sample size required to perform a test with a given power.

As evidence of the seriousness with which power analysis is now being taken, in the Editor's Preface of the very first issue (July, 2015) of the *Journal of the Economic Science Association*, the message is very clear: "A necessary (but not sufficient) condition for publishing a replication study or null result will be the presentation of power calculations."

3.1 Power Analysis - Theory

We will consider the standard situation in experimental economics, in which there are two samples, a control and a treatment, and the objective of the study is to discover whether there is a significant difference in the outcome between the two samples. Power analysis can be used to determine the sample size that is required to meet this objective.

Let μ_1 and μ_2 be the population means of the control group and the treatment group respectively. The null hypothesis of interest is $\mu_2 = \mu_1$ (i.e. the treatment has no effect), and the alternative is $\mu_2 - \mu_1 = \delta$ (i.e. the treatment has an effect of magnitude δ).[13] δ is known as the effect size and it is necessary to specify its value at the outset in order for the problem of finding the required sample size to be properly defined. The chosen value of δ is assumed to be derived either from prior beliefs, from a previous study, or from a pilot study. If none of these are feasible, Zhang and Ortmann (2013) suggest that the chosen effect size should be the smallest effect size with "economic significance".

The plan is to collect samples of size n_1 and n_2 (control and treatment respectively) for the purpose of conducting this test, and we need to decide what n_1 and n_2 should be. Recall that, before we do this, we need to set two quantities. The first is the test size, α, which is the probability of rejecting the null hypothesis when it is true (or the probability of type I error). The second is the probability of failing to reject the null hypothesis when it is false (or the probability of type II error). This second probability is conventionally labelled β. Note that the probability of rejecting the null hypothesis when it is false is $1-\beta$ and this is the *power* of the test. We shall denote power by π.

As mentioned earlier, it has become standard to set α to 0.05, unless there are compelling reasons to do otherwise. For power, many researchers use $\pi = 0.80$ as a standard for adequacy.

Having decided on these values of α and β, we proceed to apply power analysis. The test that will be performed is the independent samples t-test, which is based on the following test statistic:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.1}$$

[13] Alternative hypotheses nearly always involve inequalities, for example, $\mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$. However, in the context of power analysis, it is necessary for both the null and the alternative hypotheses to be equalities, in order for the problem to be properly defined.

where n_1 and n_2 are the sample sizes in each group, $\bar{y}_1$ and $\bar{y}_2$, are the two sample means, and s_p is the pooled standard deviation.

For the sake of simplicity, we will constrain $n_1=n_2=n$. That is, we will require that the number of observations in each treatment will be the same and given by n.[14] With this constraint, the test statistic (3.1) becomes:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{2}{n}}} \tag{3.2}$$

Under the null hypothesis, *t* defined in (3.2) has a *t(2n-2)* distribution. Hence, the rejection rule, given our chosen value of α, is $t > t_{2n-2,\alpha}$.

A complication arising here is that the rejection rule is different for every sample size. However, based on the anticipation that the value of *n* eventually chosen will be reasonably large, the normal approximation may be used and the rejection rule becomes $t > z_\alpha$, where z_α is the upper α critical value of the standard normal.[15] This simplifies the analysis considerably.

The power of the test is then given by:

$$P(t > z_\alpha | \mu_2 - \mu_1 = \delta) = P\left(\frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{2}{n}}} > z_\alpha \middle| \mu_2 - \mu_1 = \delta\right) = \Phi\left(\frac{\delta - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}}\right) \tag{3.3}$$

If the desired power of the test is 1-β, we have:

$$\frac{\delta - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}} = z_\beta \tag{3.4}$$

Rearranging (3.4) we obtain the required sample size:

$$n = \frac{2{s_p}^2(z_\alpha + z_\beta)^2}{\delta^2} \tag{3.5}$$

Once again applying our chosen values of α and β, we have z_α = 1.645 and z_β = 0.84, the formula (3.5) for the required sample size becomes:

$$n = \frac{12.37{s_p}^2}{\delta^2} \tag{3.6}$$

[14] One reason for having different sample sizes different between treatment and control is if sampling costs differ between the two. Simple rules of thumb for such situations are summarized by List et al. (2011).

[15] In Section 3.2, we verify that this approximation makes very little difference to computed power, even when n is small.

3.2 Power Analysis - Practice

Let us use (3.3) and (3.6) to conduct a power analysis on the t-test we conducted earlier on the simulated dictator game data in Section 2.6. Recall that: there were 30 subjects in each treatment; the means of giving in the no-communication and communication treatments were 13.8 and 19.4 respectively, implying a treatment effect of +5.6; the pooled standard deviation was 15.9.

Using (3.3), we find the power of this test to be:

$$P\left(t > z_{\alpha} \mid d = 5.6\right) = \Phi\left(\frac{5.6 - 1.645\left(15.9\right)\sqrt{2/30}}{\left(15.9\right)\sqrt{2/30}}\right) = \underline{0.39} \tag{3.7}$$

This clearly falls well short of the benchmark power of 0.80. The next question is what would the sample size need to be to bring the power up to 0.80. To answer this we use (3.6):

$$n = \frac{12.37\left(15.9\right)^2}{5.6^2} = \underline{99.7} \tag{3.8}$$

Since we require the power of the test to be at least 0.80, it is appropriate to round this number upwards, and conclude that the required sample size is 100 subjects in each treatment, and therefore the required total sample size is 200.

The two calculations performed above may be performed using the immediate[16] command "power" in STATA (StataCorp, 2019).[17] To obtain the power of the test, the following command is used:

power twomeans 13.8 19.4, sd(15.9) n(60) oneside

To obtain the required sample size, the n(.) option in the command is replaced by the power(.) option:

power twomeans 13.8 19.4, sd(15.9) power(0.8) oneside

Graphs of power functions can easily be obtained in STATA by using the graph option with the power command. Figure 3.1 (upper panel) presents the graph obtained when the graph option is used with the first power function above (specifying a range of values in the n(.) option). It shows power against total sample size, and shows that this relationship is concave. Figure 3.1 (lower panel) presents the graph obtained from the second power command above (with a range of values in the power(.) option). This graph shows the required sample size against the desired power, and shows that this relationship is convex. Both graphs can be used to verify that a total sample size of 200 corresponds to a power of 0.80.

[16] An "immediate" command in STATA is one which simply performs a calculation using numbers inputted to the command line, and does not use any data stored in memory.

[17] The STATA power command assumes the t-distribution and therefore provides exact solutions. Hence it can be expected to provide answers slightly different from (3.7) and (3.8) which rely on the normal approximation.

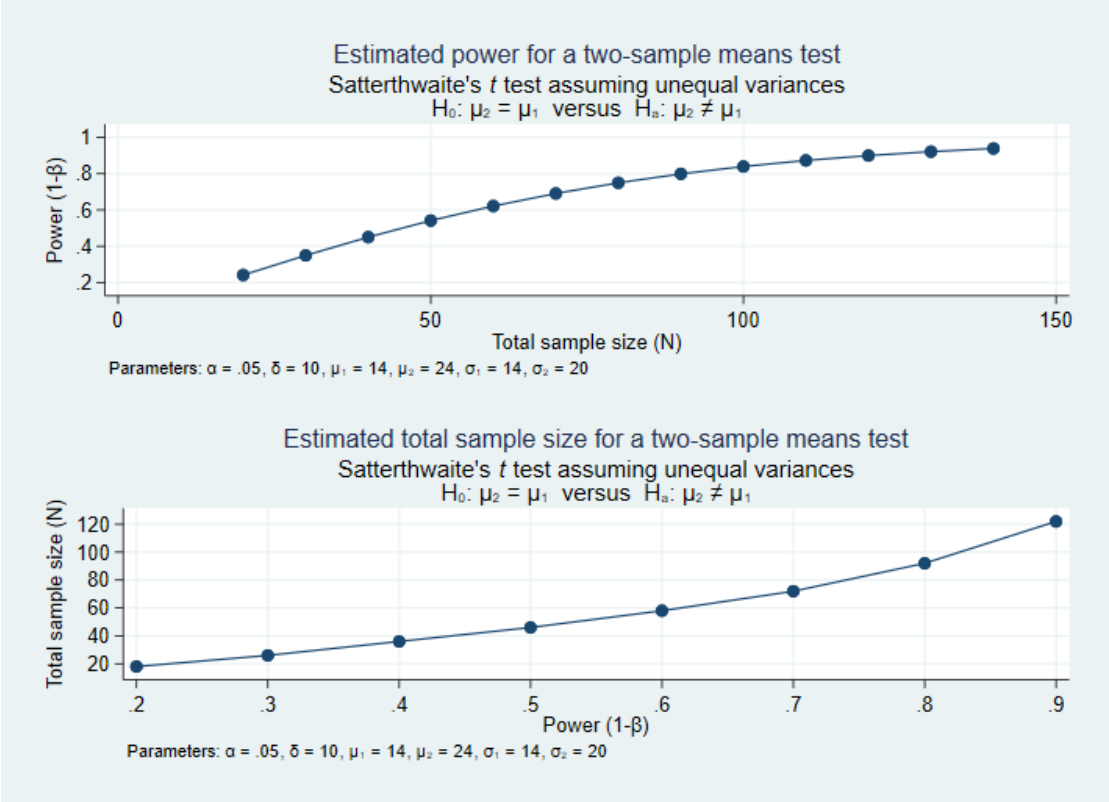

Figure 3.1: Upper panel: Power against total sample size; Lower Panel: Total required sample size against desired power.

Graphs of power functions are also useful for comparing approximate power (power obtained using the normal approximation as described in Section 3.1) with exact power (obtained in STATA which uses the exact distribution of the test statistic). This is done in Figure 3.2. We see that the approximation may be considered a touch inaccurate (i.e. slightly too high) when the sample size is very low. However, when the total sample size is greater than about 30, there is very little difference between approximate and exact power.

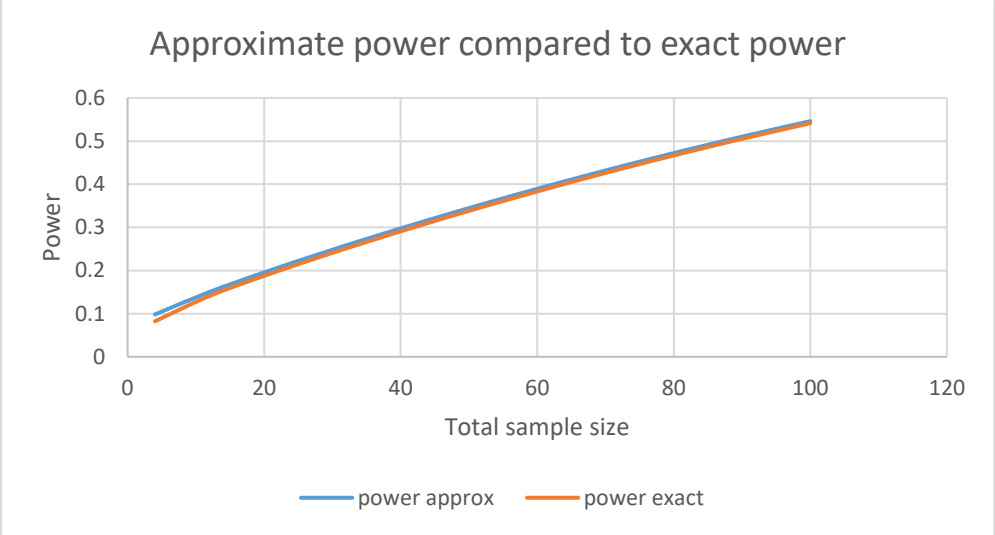

Figure 3.2: Comparison of approximate power (based on normal approximation; higher curve) with exact power (based on t-distribution), against sample size. Test parameters same as in previous examples.

The power of the paired-comparison test can also easily be found using the power command in STATA, and an important question is whether this test provides more power than the corresponding independent samples test. With the following command,

power pairedmeans 13.8 19.4, sd(15.9) n(30) oneside corr(0)

we find that the power of the paired test with 30 subjects is 0.38 – exactly the same as the power of the independent samples test with two groups of 30 subjects. However, note the option corr(0). This indicates that the correlation in giving between the two treatments is assumed to be zero. It is natural to assume that if a subject gives generously in one treatment, they are also likely to give generously in the other treatment; hence it should be assumed that the correlation is positive. In Table 2.2, we see that this correlation is +0.78. This is the number we should use in the power command:

power pairedmeans 13.8 19.4, sd(15.9) n(30) oneside corr(0.78)

When this simple change is made, we find that the power of the test rises from 0.38 to 0.88. We further find that a sample size of 24 (down from 30) would be sufficient to reach the benchmark power of 0.80.

It is easy to understand why an increase in this correlation causes a reduction in the required sample size, if we consider an extreme case. Imagine that every subject has exactly the same treatment effect, so that the correlation between the two responses is the maximum +1. If we know that all subjects have the same treatment effect, then obviously we only need to observe one subject in order to find what the treatment effect is.

The huge potential advantages of the paired-comparison approach over the independent-samples approach are made very clear by this example. We have just seen that for an independent samples test, the total number of subjects required is 200 (100 in each treatment), while for a paired-comparison test of the same hypothesis, a (total) sample of only 24 is required to achieve the same (benchmark) power.

In summary, paired tests are desirable because they allow at least a 50% saving in the number of subjects required for a given power, and the saving can be much greater than 50% in situations in which the paired responses are highly correlated.

As mentioned in Section 2, a much-discussed disadvantage of paired designs is the possibility of "order effects", that is, the behaviour of subjects depending on the order in which the treatments are experienced (see Holt and Laury, 2005). "Crossover designs" are a way of countering this problem: half of subjects see control followed by treatment; the other half see treatment followed by control. Differences between these two groups would confirm the existence of an order effect, which would then need to be controlled for in treatment tests. Also popular are "ABA designs", which start with a baseline period in which no treatment is given (A), followed by a period in which the treatment is introduced (B), and then a period in which the treatment is removed (A). This makes it possible to measure behaviour before treatment, during treatment, and once treatment is removed.

In this Section, we have illustrated the use of the very useful STATA power command as a means of performing power calculations applying to between-subject ("twomeans") and within-subject ("pairedmeans") t-tests. It may be useful to mention at this stage that there

are many other applications of this command in STATA, including the chi-squared test, Fisher's exact test, and the t-test for a slope in a regression.[18]

3.3 Power and the scientific quality debate

There has recently been fierce debate in the Experimental Economics literature on the subject of scientific quality. See for example Camerer et al. (2016). The central question is under what circumstances and to what extent can published experimental results be trusted. The concept of power in tests plays a key role in this debate.

Low power appears to be common in Experimental Economics. Zhang and Ortmann (2013) report that average power in a sample of dictator game experiments is only 0.38. Although this statistic may be perceived as an embarassment to Experimental Economists, it does not seem too bad when placed in perspective. Ioannidis et al. (2018) surveyed 6,700 studies spread over 159 different topics from Empirical Economics generally, and found that median power is only 0.18. Power of a typical published study is clearly woefully lower than the received benchmark of 0.80.

Zhang and Ortmann (2013) emphasise the importance of calculating a sample size *ex ante* which is powerful enough to detect an effect size of *economic significance* (i.e., an effect size that reflects the implied economic benefits and costs). The importance of choosing the sample size ex ante is that failing to do this leaves scope for the practice of continuing to collect new data until a statistically significant effect is found. It can be demonstrated that this practice greatly inflates the probability of type I error.[19] This is one major cause of "false positive" results in experimental economics.

Another important cause of false positives is what is known as the "desk-drawer effect", which is closely related to the problem of publication bias. Consider a study that fails to find a significant treatment effect as a consequence of low power. Such a study is less likely to be written up, less likely to be submitted for publication, and less likely to be accepted for publication. By the same token, the under-powered studies that reach publication tend to be the ones that find a high effect size by chance. Hence, for both of these reasons, there is a tendency for published effect sizes to be exaggerations of true effect sizes.

This has important implications for the replicability debate. If a replication study is planned, and a power analysis is performed on the basis of the effect size obtained in the original published study, and if that effect size is biased upwards for the reasons suggested in the last paragraph, the replication study will inevitably be under-powered.

Camerer et al. (2016) make two specific recommendations to address these problems; pre-registration; and registration and reporting of all study results. Pre-registration of analysis plans is likely to reduce the over-incidence of false positives in initial experiments. Registration of study results is likely to reduce the effects of publication bias on effect sizes.

3.4 The Monte Carlo Method in power calculations

[18] To obtain a complete list of applications of the power command in STATA (StataCorp, 2019), select "Statistics" from the toolbar and then select "Power, precision and sample size" from the dropdown menu.

[19] Kruschke (2011) uses the analogy of coin flips. If one flips a coin up to 20 times, and stops each time to test the hypothesis that the coin is fair, the false rejection rate is 17.1% rather than 5%.

The STATA power command, demonstrated in previous sections, is very useful for carrying out power calculations for simple parametric tests, particularly tests based on the t-distribution. If we require to find the power of a parametric treatment test in a complex model, or the power of a non-parametric test, the Monte Carlo method is often required. The Monte Carlo method has been previously used by Forsythe et al. (1994) to compare the performance of a range of treatment tests.

Here, we will use the following simple data generating process:

$$\begin{aligned} y_i &= 10 + \delta d_i + \varepsilon_i \quad i = 1, \cdots, n \\ d_i &= 0 \; if \; i \le n/2 \\ d_i &= 1 \; if \; i > n/2 \\ V(\varepsilon_i) &= 1 \end{aligned} \tag{3.9}$$

In (3.9), y_i is the decision variable which is assumed to have a mean of 10 under the control. d_i is a dummy variable representing treatment: 1 for treatment; 0 for control. The first half of the sample is control; the second half treatment. The total sample size is n. The parameter δ is the treatment effect.

We set the sample size to n=100, so that there are 50 in each treatment. We assume three different distributions for ε_i: normal; uniform; and skewed ($\chi^2(3)$). All distributions are standardised to have mean zero and variance 1.

We use a simulation with 10,000 replications to find the actual size of each test – that is, the proportion of replications for which the test rejects the null hypothesis of no effect when there is no effect. We then use a fresh simulation with 10,000 replications to find the power of each test when the true effect size is δ=0.5 – that is, the proportion of replications in which the null hypothesis is correctly rejected.

	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

Table 3.1. Monte Carlo estimates of size and power of four tests: t-test; Mann-Whitney (MW); Kolmogorov-Smirnoff (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.

The results are reported in Table 3.1. Firstly, note that the power of 0.692 estimated for the t-test applied to normally distributed data is a power that can be obtained using the power command in STATA. For all of the other powers in the Table, the Monte Carlo method is required.

From Table 3.1, we see that all tests are close to being correctly sized, with the exception of the KS test which is possibly undersized.[20] Which test has the best power performance depends crucially on the distribution of the data. With normally distributed data, the t-test performs best with the Mann-Whitney test a close second. With uniform data, the Epps-Singleton test appears to overtake the others in performance. With skewed data, the original order appears to be completely reversed, the Epps-Singleton test performing best (with a power close to 1) and the t-test performing worst.

The clear message here is that the distribution of the data is very important in the choice of treatment test. As theory informs us, the t-test can be relied upon if the data is normally distributed. However, if the data is non-normal, and particularly if the distribution of the data is asymmetric (as is very common in the experimental setting) other tests are likely to perform better. Of particular interest is the strong performance of the Epps-Singleton test in conditions of non-normality, and this is an important issue for further investigation.

3.5 Power of Treatment tests in multilevel models

As discussed in Section 2.15, it is unusual for experimental data to consist of independent observations. There is often dependence at different levels. For example, if each subject makes a sequence of decisions, there is dependence at the level of individual subjects. In interactive experiments, there is also likely to be dependence at the level of the group of subjects, or at the level of the session in which the groups of subjects perform their tasks.

Many methods are available for dealing with the complicated structure of an experimental data set. In this section we use Monte Carlo analysis in an attempt to assess how useful some of such methods are. For example, how serious is it to ignore the clustering in the data, and just proceed with OLS and OLS standard errors? Which model performs best under the

[20] An "under-sized" test is one whose actual size is less than nominal size, meaning that the frequency of type I error is lower than it should be in theory. The reason why this is considered a problem is because size is inseparably related to power: low size implies low power. An extreme example is useful in illustrating this: a test that never rejects has an actual size of zero, but also has zero power implying that it is useless as a test.

complicated structure? The use of Monte Carlo to address such questions has been made previously by Cameron et al. (2008).

The data generating process will contain both subject level and group-level clustering. We are once again interested in tests of a treatment effect. Seven different testing procedures are used. Each is a t-test from a particular regression model. The seven models are:

- OLS no clustering
- OLS with clustering at the subject level
- OLS with clustering at the group level
- Random effects, no clustering
- Random effects, with clustering at the subject level
- Random effects, with clustering at the group level
- Multi-level model (subject random effect and group random effect)

We will consider both between-subject and within-subject tests. In this context: "between-subject" means applying the treatment to half of the subjects; "within-subject" means applying the treatment to half of the tasks.

The questions we set out to answer with the Monte Carlo are: Which of these testing methods are correctly sized? Of those which are correctly sized, which has highest power?

The data generating process is a three-level model of the form (2.8) in Section 2.14.

$$\begin{aligned} &y_{ijt} = \beta' x_{it} + \delta d_i + u_i + v_j + \varepsilon_{ijt} \\ &i = 1,\cdots,n \;\; j = 1,\cdots,J \;\; t = 1,\cdots,T \\ &V(u_i) = \sigma_u^2; V(v_j) = \sigma_v^2; V(\varepsilon_{ijt}) = \sigma_\varepsilon^2 \end{aligned} \tag{3.10}$$

In (3.10) i represents subject, t represents task, and j represents group. The treatment test of interest is a t-test applied to the treatment effect δ. The treatment dummy d_i has only an i subscript. This implies that the treatment is being applied between-subject: some subjects are exposed to the treatment throughout the experiment; others are not. Of course, it would be possible to apply the treatment within-subject, that is, for all subjects to experience the treatment for (say) half of the tasks. In this case, the treatment variable appearing in (3.10) would be d_{it}, that is, it would have both i and t subscripts.

We assume n=40 subjects and T=50 rounds. We also assume that the 40 subjects are divided into J=10 groups of 4. We use 100 replications in the Monte Carlo.

	Between-subject test		Within-subject test	
	Size	Power(δ=0.5)	Size	Power(δ=0.05)
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

Table 3.2: Results from Monte Carlo experiment with DGP (3.10). All tests have nominal size 0.05.

Results from the Monte Carlo are presented in Table 3.2. For the between-subject tests, we see that only three of the seven testing procedures result in tests that are correctly sized. Some are seriously over-sized. Most spectacularly, the test size under ols without clustering is 0.46. This means that when clustering is completely ignored, a significant treatment effect is found nearly half of the time, *even though the true effect of the treatment is zero*. Hence the importance of dealing with clustering is clear.

Interestingly, the clustering models that result in unbiased tests (i.e. with actual size significantly different from nominal size) are the ones that deal with clustering at the group level. Dealing with clustering at the lower level of the individual appears to be inadequate.

When deciding which of the testing procedures is best, we restrict attention to the three unbiased (i.e. correctly-sized) procedures, and then look at power. We see that of the three, the multi-level model gives the highest power of 0.27 (although the difference in power between the three is not great). On this basis, we may conclude that the multi-level model is the best framework in which to conduct the between-subject treatment test.

We next turn to the within-subject tests, in which all subjects experience the treatment in half of the rounds. Since within-subject tests can detect smaller treatment effects, we shall assume a much smaller treatment effect of δ=0.05 under the alternative hypothesis. We see very different results from the between-subject tests. All seven tests appear to be unbiased. The testing procedure that ignores clustering has very low power. The other six have modest power, and there is very little difference between them.

The key results of this section summarised as follows. In the between-subject context, only three of the seven tests are correctly sized: the two that use group-level clustering, and the multi-level model. Of these three tests, the most powerful is the one performed in the framework of the multi-level modelling. Failure to deal with clustering has very serious consequences in terms of massively excessive test size. In the within-subject context, the results are very different. Firstly, within-subject tests are able to detect much smaller treatments than within-subject tests. Secondly , all of the approaches perform well on both size and power, except OLS without clustering.

Recommendations that follow from these results are as follows. In the between-subject context, the multi-level model is the best model in which to conduct treatment tests. If clustering is to be used, it is preferable to cluster at the highest possible level (e.g. group rather than subject). In the within-subject context, all that appears to matter is that OLS without clustering is avoided.

3.6 Choosing number of subjects and number of tasks

Power analysis provides a framework for deciding on the required number of subjects (n) for an experiment. But since most experiments require subjects to engage in a sequence of tasks, we also need to consider what is the appropriate number of tasks (T). Related questions follow. For example, given that the experimenter is working subject to a budget constraint, it is plausible that they might face a trade-off between n and T. For example, given an increase in the budget, is it more beneficial to increase the number of subjects, or to increase the number of tasks (and to provide higher incentives to compensate)? Or, should one of n and T

be increased at the expense of the other? Although List et al. (2011) provide useful rules of thumb that relate to these questions, it certainly seems to be un under-researched area.

Since in the last section the multi-level model was established as the best framework in which to conduct a treatment test, in this section we shall restrict attention to this model, and investigate the effect of varying n and T on power.

The DGP is again based on (3.10), and we consider both a between-test and a within-test. n takes values of 40, 80 and 120, and T takes values of 50, 100, 150. Power for each test and for each combination of n and T are presented in Table 3.3.

	Between-subject test (δ=0.50)			Within-subject test (δ=0.05)		
	T=50	T=100	T=150	T=50	T=100	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

Table 3.3: 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).

For between-subject tests, we see that increases in n and T do tend to bring about increases in power, but these increases to not appear to be very steep. In fact, if we go on increasing both n and T, power seems to level off at a ``power ceiling'' of around 0.40.

Again we see results that are very different in the within-subject setting. Aside from the ability of the within-sample test to detect the much smaller treatment effect, it seems that increases in n and T both bring about steep increases in the power of the test. At the highest values of n and T considered, power is almost 1. Notice also that increases in T appear to be slightly more beneficial than increases in n.

4. Experimental Data types

A variety of data types present themselves in Experimental Economics. These include binary data, ordinal data, interval data, exact data, and censored data. We shall consider each of these. The context in which each of these data types are introduced is principally the measurement of risk attitude. This is a convenient context because many different data types can be used for this purpose.

For most of this section, we restrict attention to the situation of a single observation per subject. In a later section, we will consider models suitable for settings in which there are multiple observations per subject.

4.1 Binary data

There are many examples in experimental economics of binary data models (usually logit or probit) being used in the estimation of preference parameters. A very popular example is the modelling of behaviour under risk. Binary data on choices between pairs of lotteries are used

to estimate subjects' levels of risk aversion, and in some cases also probability weighting parameters. See, for example, Hey and Orme (1994), Loomes et al. (2002), Harrison and Rutstrom (2009), and Conte et al. (2011). Another popular example is the analysis of intertemporal choice: subjects choose between different money amounts to be received at different future dates. Such data can be used to estimate subjects' individual discount rates, and to test for hyperbolic discounting versus exponential discounting (see Andersen at al., 2008). A third example is the analysis of the recipient's decision in the ultimatum game. In this game the recipient makes a binary decision on whether to accept or reject the proposer's allocation of a pie. Analysis of this decision can be used to model the dependence of propensity to accept an offer on the amount of the offer, and conclusions can be drawn about subjects' aversion to disadvantageous inequality (see Roth et al., 1991).

Moffatt (2015) considers a very simple setting in which subjects choose between two lotteries S and R, where:

S: \$5 with certainty
R: (0.5, \$0; 0.5, \$10)

S is the "safe" lottery and it pays \$5 with certainty. R is the "risky" lottery and represents a 50:50 gamble involving the outcomes \$0 and \$10. Note that both lotteries have the same expected value of \$5, and hence, assuming expected utility (EU) maximisation, choice of S implies risk aversion, choice of R implies risk seeking, and indifference between S and R implies risk neutrality. Before making the choice, each subject was endowed with an amount \$w (wealth), and the focus of the investigation is the effect of w on risk attitude. The hypothesis under test is the "house-money effect" (Thaler and Johnson, 1990; Keasey and Moon, 1996), according to which behaviour becomes less risk-averse as the wealth endowment increases.

Popular models that are useful for the analysis of this sort of experimental data are binary probit and binary logit. Binary probit is defined by:

$$P(y_i = 1 | w_i) = \Phi(\beta_0 + \beta_1 w_i) \quad i = 1, \cdots, n \tag{4.1}$$

where y_i is the binary dependent variable (1 if subject i chooses S, 0 if R), and $\Phi(.)$ is the standard normal cdf. Binary logit is defined by:

$$P(y_i = 1 | w_i) = \Lambda(\beta_0 + \beta_1 w_i) \quad i = 1, \cdots, n \tag{4.2}$$

where $\Lambda(.)$ is the logistic function.

The parameters of (4.1) and (4.2) are typically estimated by maximum likelihood. The results provide a direct test of the house money effect: a significantly negative estimate of the coefficient of w provides strong evidence that an increase in endowment causes a fall in the probability of choosing S, that is, evidence in favour of the house money effect.

In the context of this example, Moffatt (2015) also demonstrates the use of the margins command in STATA 16 (StataCorp, 2019) to obtain marginal effects and predicted probabilities following estimation of these models. He also demonstrates the use of the delta method (Oehlert, 1992) to deduce an estimate of the wealth level that induces risk-neutrality.[21]

[21] The wealth level that induces risk neutrality in the context of models (4.1) and (4.2) is given by

4.2 Optimal design of binary choice problems

An interesting problem from the viewpoint of an experimental economist is how a set of risky choice problems (or other types of choice problem) should be designed (i.e. how the probabilities and outcomes defining each choice problem should be decided) in order to optimise with respect to the efficiency in estimation of the parameters of interest. This is another problem in experimental design. However, to solve it requires a very different approach to that of power analysis which was used to determine the required sample size.

In the statistical literature, the concept of an optimal experimental design is frequently taken to mean a "D-optimal design"; that is, one that maximises the determinant of Fisher's information matrix (see e.g. Atkinson, 1996). Maximising this quantity is equivalent to minimising the volume of the "confidence ellipsoid" surrounding the point estimates, and is hence equivalent to maximising estimator precision.

In a linear regression context with a single independent variable, the solution to this problem takes the trivial form of choosing values of the independent variable that occupy the "corners of the design space". That is, there are only two design points: the lowest and highest permissible values of the independent variable.

The binary data setting is more complicated. The information matrix for the probit and logit models, defined in the last Section, both have the same form, which is:

$$I = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} \tag{4.3}$$

For the probit model, w_i in (4.3) is given by:

$$w_i = \frac{\left[\phi\left(\beta_0 + \beta_1 x_i\right)\right]^2}{\Phi\left(\beta_0 + \beta_1 x_i\right)\left[1 - \Phi\left(\beta_0 + \beta_1 x_i\right)\right]} \tag{4.4}$$

For the logit model:

$$w_i = \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{\left[1 + \exp\left(\beta_0 + \beta_1 x_i\right)\right]^2} \tag{4.5}$$

For both models, the determinant of the information matrix (4.3) may be written as (see Kanninen, 1993):

$$\det(I) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_i w_j \left(x_i - x_j\right)^2 \tag{4.6}$$

The optimal design problem amounts to choosing values of *x* to maximise (4.6). As with the linear model, there are only two design points. However, in (4.6) we see that det(I) is weighted by the w_i's. These weights are maximised when (in the probit case) $\Phi\left(\beta_0 + \beta_1 x_i\right) = 0.5$, that is, when the probabilities of the two outcomes are equalised. So, the objective of having design points as far from each other as possible is countered by the objective to have design

---

$-\beta_0/\beta_1$. The STATA command for applying the delta method is nlcom. The major advantage from using the delta method is that it provides a standard error and confidence interval for the quantity being estimated.

points giving rise to perfect indifference. The latter is the requirement of "utility balance" (Huber and Zwerina, 1996).

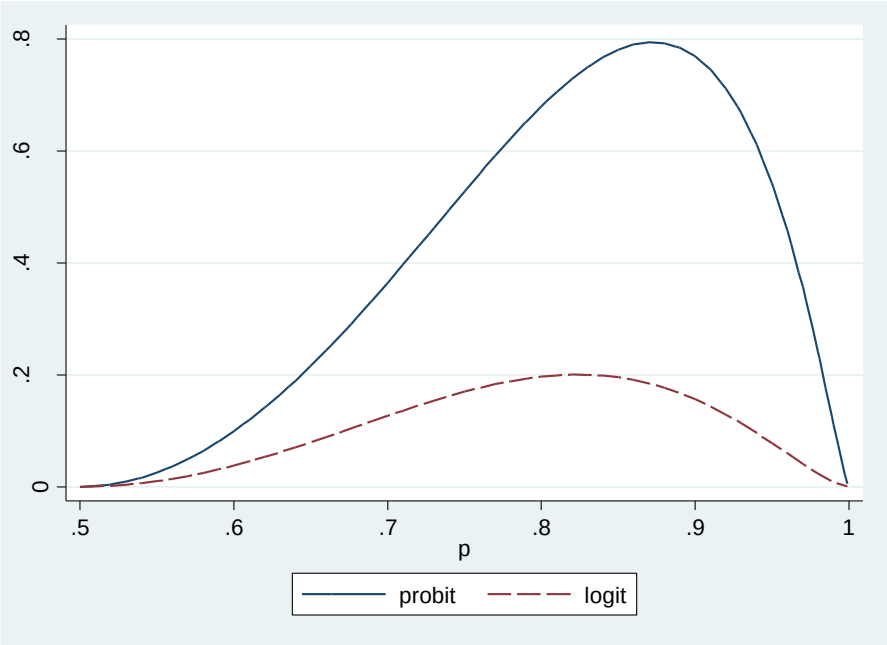

Figure 4.1: Determinant of information matrix against percentile of larger design point; probit and logit.

In Figure 4.1, det(I), given in (4.6), is plotted against the percentile of the upper design point, for both probit and logit. The design will be symmetric, so the lower design point will be an equal distance from the centre. We see that when both design points are in the centre (percentile = 0.5), the information is zero. The intuition for this is that if the tasks are designed so that subjects are exactly indifferent, the spread of the distribution will not be identifiable. We also see that when both design points are the maximum distance from the centre (percentile = 1), the information is again zero. The intuition for this is that when subjects are presented with problems for which their responses can be predicted with certainty, the data will be of no value. The most important features of Figure 4.1 are that, for the probit model, information is maximised at 0.87, implying that the optimal design points are at the 13th and 87th percentiles, while for the logit model, information is maximised at 0.82, implying that the optimal design points are at the 18th and 82nd percentiles.

It is perhaps surprising that the optimal design points for probit are 5 percentage points further into the tails than those for logit, given that we are often led to believe that the two models are very similar.[22]

An important point is that, in order to apply these optimal design rules, the values of the parameters of interest must be known. Moffatt (2015) assumes a Random Preference model[23] with EU, in which that the power utility parameter, r, has the following distribution for a typical subject:

$$\ln(r) \sim N(-0.89, 0.16^2) \qquad (4.7)$$

[22] For example, according to Greene (2008, p.774), "in most applications the choice between [probit and logit] seems not to make much difference".

[23] See Section 6.9 for a full explanation of the Random Preference Model.

If the probit model is to be used for estimation, the two optimal design points are at the 13th and 87th percentiles of this distribution, which correspond to r-values of 0.34 and 0.49. All that remains is to reverse-engineer choice problems which have these threshold r-values.[24] Of course, for any given threshold r-value, there are an infinity of possible choice problems. Hence Moffatt restricts these possibilities by requiring that the safer outcome is a certainty of 10, and the riskier involves only the outcomes 0 and 20. For a threshold r-value of 0.34, the probability of the higher outcome (20) in the riskier lottery must be 0.79, while for the threshold value of 0.49, the said probability is 0.70. Thus two optimal choice problems are created for the subject whose preferences are defined by (4.7).

Subjects to not all have identical prefrences, and subject heterogeneity may be introduced by allowing each subject to have a different mean in (4.7). A further assumption therefore needs to be made about the between-subject distribution of this mean. Since no prior information is available on the risk attitude of individual subjects, Moffatt (2015) recommends a strategy of deriving the optimal choice problems, in the way outlined above, over a range of subject preferences (i.e. over a grid of values of the mean in (4.7)). See Moffatt (2015) for further details, and for the resulting optimal design of a risky choice experiment.

While the design of Moffatt (2015) is optimal for the objective of estimating the within-subject and between-subject distributions of risk attitude under EU, Moffatt (2007) goes further by considering optimal designs where the objective is the estimation of probability weighting parameters (i.e. allowing departures from EU).

A problem that may have become apparent in the reading of this section is that, in non-linear models such as those considered, the parameters of the model need to bo known in advance in order find the D-optimal design. This presents a logical problem:[25] an experiment cannot be designed without knowledge of the parameters whose values the experiment's purpose is to discover! Apart from the obvious solution of using parameter estimates from a previous study or pilot study, a possible remedy is a sequential design, in which choice problems are tailored to individual subjects, using their choices in early problems to identify their risk attitude, and then set later problems using an optimal design rule for that risk attitude. The obvious drawback of this approach is the potential violation of incentive compatibility: subjects may manipulate the experiment by deliberately making false responses in an effort to steer the problem sequence in the direction of the most desirable problem types. Johnson et al. (2019) have developed a sequential design in such a way as to meet the requirement of theoretical incentive compatibility.

4.3 Ordinal Data

Ordinal data can be seen as a generalisation of binary data to a situation in which the dependent variable can take more than two possible values. Consider the examples of binary data used at the start of Section 4.1. Let us suppose that, having chosen between the two options in the choice problem, subjects are then asked how sure they are about their choice: "not sure", "fairly sure", or "completely sure". This gives a set of six possible ordered outcomes, shown in Table 4.1. The variable containing the ordered outcome is named ys. The variable ys may be seen to represent "strength of preference", a concept that has been studied in Experimental Economics by Connolly and Butler (2006) and Butler et al. (2014). The

[24] The threshold r-value for a choice problem is the value of r that makes a subject indifferent between the riskier and safer choices.

[25] This problem is often referred to as the "chicken-and-egg" problem.

measurement scheme used to obtain variables such as ys is often referred to as a Likert scale (Likert, 1932).

Binary choice	*y=0*	*y=0*	*y=0*	*y=1*	*y=1*	*y=1*
Sureness	Completely sure	Fairly sure	Not sure	Not sure	Fairly sure	Completely sure
Ordered outcome	*ys=1*	*ys=2*	*ys=3*	*ys=4*	*ys=5*	*ys=6*

Table 4.1: Obtaining an ordinal variable from a binary choice variable.

To model ys, two very useful estimation frameworks are ordered probit and ordered logit (Daykin and Moffatt, 2002; Greene and Hensher, 2010). These models assume an underlying latent variable $y_i^*(-\infty < y_i^* < \infty)$ representing respondent i's propensity to respond positively. The variable y_i^* may be perceived as the dependent variable in a linear regression model. However, y_i^* is not fully observed. All that is known about y_i^* is that it falls in an interval corresponding to the observed ordinal response.

The inderlying latent variable is assumed to depend linearly on a set of independent variables contained in the vector x_i, as follows:

$$\begin{aligned} &y_i^* = x_i'\beta + u_i \quad i = 1,\cdots,n \\ &u_i \sim N(0,1) \end{aligned} \tag{4.8}$$

β is a vector of parameters**,** *not* containing an intercept. The assumption of normality of the error term in (4.8) is for ordered probit; for ordered logit, the error term is assumed to be logistic. The relationship between the latent variable *y** and the observed variable *y* is:

$$\begin{aligned} &y=1 \text{ if } -\infty < y^* < \kappa_1 \\ &y=2 \text{ if } \kappa_1 < y^* < \kappa_2 \\ &y=3 \text{ if } \kappa_2 < y^* < \kappa_3 \\ &: \\ &y=J \text{ if } \kappa_{J-1} < y^* < \infty \end{aligned} \tag{4.9}$$

The parameters κ_j, *j=1,…,J-1*, are known as the "cut-point" parameters.

Based on a sample (y_i, x_i, *i=1,…,n*), the log-likelihood function is (for ordered probit):

$$LogL = \sum_{i=1}^{n} \ln\left[P_i(y_i)\right] = \sum_{i=1}^{n} \ln\left[\Phi\left(\kappa_{y_i} - x_i'\beta\right) - \Phi\left(\kappa_{y_i-1} - x_i'\beta\right)\right] \tag{4.10}$$

where $\Phi(.)$ is the standard normal cumulative distribution function. The log-likelihood (4.10) is maximised with respect to the elements of β along with the cut-points $\kappa_1, \kappa_2, \ldots, \kappa_{J-1}$, to give maximum likelihood estimates (MLEs) of both sets of parameters.

Note that when there are only two possible outcomes on the Likert scale, for example agree/disagree, ordered probit (resp. ordered logit) simplifies to the more familiar binary probit model (resp. binary logit) covered in Section 4.2 with the only difference that the single cut-point, although equal in magnitude to the intercept arising from binary probit, is opposite in sign.

It is useful to compare the regression parameters estimated using the ordered probit model with sureness as the dependent variable, to those estimated using the binary probit model with the choice as the dependent variable. The former estimate is clearly more efficient since ordinal data embodies more information than binary data. However, there is a reason to be wary when modelling ordinal outcomes. This is that asking subjects for strength of preference cannot be incentive compatible, and hence, in the view of many economists, the ordinal data analysed above cannot be trusted.[26] In contrast, the straightforward binary choice between the two lotteries is - assuming that the prize will be paid out – incentive compatible.

These considerations lead to the question of whether there is a way of testing consistency between the ordinal data and the binary data, in order to judge whether the ordinal data is reliable. Moffatt (2015) has argued that the Hausman specification test (Hausman, 1978) is suitable for this purpose. The null and alternative hypotheses are:

H_0: Subjects respond truthfully when asked for their strength of preference.
H_1: Subjects do not respond truthfully when asked for their strength of preference.

The intuition of the Hausman test is as follows. Estimation of the preference parameters using binary choice data may be assumed to be consistent, since the binary choice is incentive compatible. However, this estimator is not efficient because it does not make use of all available information. Estimation using the ordinal data is both consistent and efficient, *provided H_0 is true*. If H_0 is false, this estimator is inconsistent. Hence the two estimators satisfy the conditions required for the Hausman test. The Hausman test statistic is based on the difference between the two sets of estimates, with a large difference representing evidence against H_0.

Moffatt (2015) provides Monte Carlo evidence of the power of the Hausman Test when applied in this setting. With a sample size of 1000, and with 30% of the sample mis-reporting their strength of preference, the probability of the Hausman test correctly detecting the departure from truthful reporting is found to be 0.714.

Another area of experimental economics in which ordinal data arises is in the elicitation of emotions. Bosman and van Winden (2002) measure subjects' emotions in the course of a "power-to-take" game. This game involves a "taker" and a "responder". The essence of the game is that the taker is free to take as much of the responders endowment as she desires, while the responder is free to retaliate by reducing the amount that will be taken, at the cost of reducing his own final payoff. Clearly the emotions stirred in the respondent are highly important in this setting. On completion of the game, emotions such as anger, irritation, and surprise, are elicited on a seven-point Likert scale (Likert, 1932). The ordered logit model reveals that the take-rate (proportion taken by taker) has a significantly positive effect on some of these emotions.

Once again there is a problem of incentive compatibility: there is no logical reason for a self-interested subject to reveal his or her emotions truthfully. However, as with data on strength of preference, there is a way of assessing the extent of this problem. The emotion data can be compared to observed behaviour in the game. For example, two of the emotions which are stirred by high take-rates (irritation and contempt) are the same two emotions that bring

[26] According to some economists, results from economic experiments cannot be taken seriously in the absence of task-related incentives. See, for example, Grether and Plott (1979). Consider also the well-known "precepts" of Smith (1982).

about an increase in the destroy rate. This agreement is, to an extent, consistent with the validity of the emotion data.

4.4 Interval Data

Interval data arises when a variable is not exactly observed, but each observation is known to fall in a particular range. When a dependent variable has this feature, the appropriate model is the interval regression model. This model is applied in many areas of economics, such as models of income determination (see e.g. Daniels and Rospabé, 2005).

Interval data is similar to ordinal data, with the important difference that with interval data the positions of the cut-points are known, and hence they do not need to be estimated. Some researchers have been known to apply the ordered probit model to interval data; this is, of course, inefficient because it involves the unnecessary estimation of the known cut-points. Other researchers have been known to convert the observations of the dependent variable into midpoints of the interval in which the observation is known to lie, and then proceed with linear regression. The resulting estimator is known to be inconsistent (Stewart, 1983).

Interval data arises when a multiple price list (MPL) has been used for the elicitation of risk attitude.[27] A very popular MPL is the one designed by Holt and Laury (2002). The Holt-Laury MPL is presented in Table 4.2. There are 10 choice problems listed in order, in each of which the subject is required to choose between a safe lottery (S) and a risky lottery (R). In Problem 1, we expect all subjects to choose S; in Problem 10, we expect all subjects to choose R (in fact, R stochastically dominates S in Problem 10). What is interesting is where in the sequence a subject switches from S to R, since this will indicate their attitude to risk.[28]

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)	∞

Table 4.2: The Holt and Laury (2002) multiple price list, with threshold risk aversion parameter for each choice problem.

We will assume that subjects have the constant relative risk aversion (CRRA) utility function (see Section 6.1 for a more detailed explanation of this and related concepts):

[27] The advantages and disadvantages of various types of MPL when used as elicitation devices are discussed by Andersen et al. (2006).

[28] A potential complication with MPLs such as that shown in Table 4.2 is the possibility that subjects may switch more than once. This presents an obstacle to the analysis of the data. An obvious solution to this problem is that subjects are induced to report a single switch-point (i.e. switching back and forth is not permitted). If the investigator is interested in within-subject variability, a different sort of design should be used, in which the choice problems are presented to subjects one at a time and in a random order. The analysis of data resulting from this sort of design is in fact the subject of a later section of the monograph.

$$U(x)=\frac{x^{1-r}}{1-r} \quad r\neq 1 \tag{4.11}$$

In (4.11) r is the coefficient of relative risk aversion. We shall assume that each subject has their own coefficient of relative risk aversion, *r*, and for this reason we shall refer to the model as the "heterogeneous agent model". We only need to make an assumption about how *r* varies over the population. An obvious choice is:

$$r \sim N\left(\mu,\sigma^2\right) \tag{4.12}$$

In the fourth column of Table 4.2, a value *r** is shown. This is known as the "threshold risk attitude" for the choice problem. It is the risk attitude (i.e. the coefficient of relative risk aversion) that would make a subject indifferent between S and R for the choice problem, assuming that subjects maximise expected utility (EU). As an example, if a subject chooses S on problems 1-3, and chooses R on problems 4-10, they are revealing that their risk attitude (*r*) is somewhere between -0.49 and -0.15.

When each subject reveals their switch-point, the sort of data that results is known as "interval data". The method appropriate for estimating the parameters of (4.12) is the interval regression model. For each subject, *i*, we have a lower bound (l_i) and an upper bound (u_i) for their *r*-value, taken from the final column of Table 4.2. The sample log-likelihood for the interval regression model is:

$$LogL=\sum_{i=1}^{n}\ln\left[\Phi\left(\frac{u_i-\mu}{\sigma}\right)-\Phi\left(\frac{l_i-\mu}{\sigma}\right)\right] \tag{4.13}$$

(4.13) is maximized to give MLEs of μ and σ. This is a regression model with no explanatory variables. If explanatory variables (such as gender and age of the subject) are included, then the mean parameter is allowed to depend on these variables, and their coefficients reveal the effect of these characteristics on risk aversion.

A similar application of interval regression was implemented by Bacon et al. (2020). They use data from the German Socio-Economic Panel (SOEP, Goebel et al., 2019). As part of this survey, respondents were asked the following hypothetical lottery investment question.

> *Imagine that you have won €100,000 in the lottery. Immediately after receiving your winnings you receive the following offer: You have the chance to double your money. But it is equally possible that you will lose half the amount invested. You can participate by staking all or part of your €100,000 on the lottery, or choose not to participate at all. What portion of your lottery winnings are you prepared to stake on this financially risky, yet potentially lucrative lottery investment?*
>
> *€100,000 (i.e. all of it);*
> *€80,000;*
> *€60,000;*
> *€40,000;*
> *€20,000;*
> *Nothing: I would decline the offer*

As with the Holt-Laury MPL, and again assuming EU, an individual's response to this question may be taken to imply that that individual's coefficient of risk aversion lies in a particular interval. Consequently, the interval regression model is again used for the analysis of this outcome.

Because the SOEP is conducted at different time periods, it is possible to include time-varying explanatory variables in this interval regression model. The hypothetical lottery question was used in two different years: 2004 and 2009. Conveniently, these two years lie either side of the global financial crisis. Distributions of the responses in the two years are shown in Figure 4.2. We see that in both years the responses are skewed towards the most risk averse choice, but that this skewness is more pronounced in 2009, suggesting that the crisis had the effect of increasing risk aversion.

One of the independent variables used by Bacon et al. (2020) is the VDAX, the standard measure of volatility in the German stock market. VDAX may be interpreted as a measure of "background risk", and is seen to be much higher, on average, in 2009 than in 2004. A positive coefficient on this variable may be interpreted as evidence of "risk vulnerability" (Gollier and Pratt, 1996). Bacon et al. (2020) indeed find strong evidence of risk vulnerability, and also obtain an estimate of the *coefficient of risk vulnerability*: a doubling of background risk causes in increase in the coefficient of absolute risk aversion of between 1.03 and 1.27.

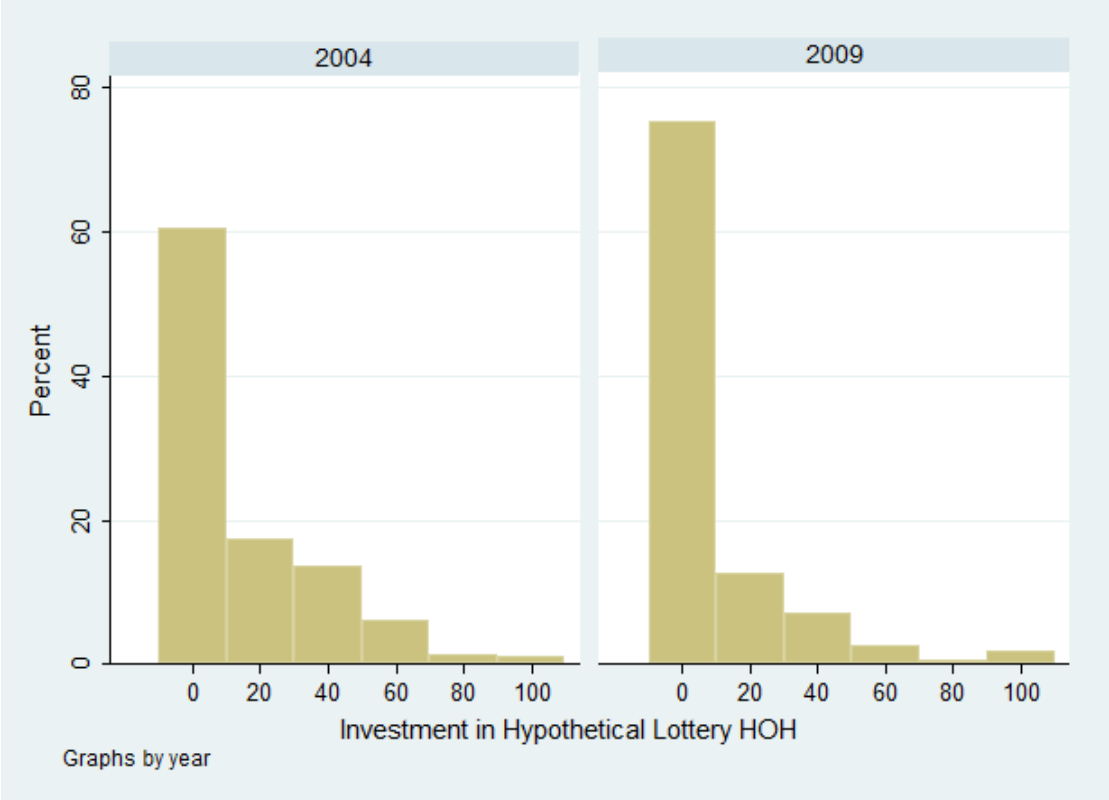

Figure 4.2: Distribution of responses to SOEP's hypothetical lottery question in 2004 (left panel) and 2009 (right panel).

4.5 Multivariate Interval Data

In order to estimate the distribution of risk attitudes over the population using MPL data, as discussed in the last section, it is necessary to assume EU maximisation. However, if subjects are given two different MPLs and asked for their switch point on each, it becomes possible to relax the assumption of EU, and to estimate the distribution of both risk attitude and probability weighting.[29] To model the resulting data, it is necessary to generalise the interval regression model to two dimensions.

Bivariate interval data is encountered in areas unrelated to experimental economics. It is useful for motivational purposes to consider a particular example. Sun and Ding (2019) consider the progression of a bilateral eye disease, and are interested in the timing of the onset of the disease. Two eyes from the same patient are perodically examined for the presence of the disease. Let t_1 and t_2 be the timing of the onset of the disease in the left and right eyes respectively. With reference to Figure 4.3, let us suppose that patients are examined after 1 year, 2 years and 3 years. For Patient 1 the disease is detected after 2 years in the left eye and after three years in the right. Since examinations are only administered once per year, the exact time of onset is unknown. Hence, for Patient 1, the intervals $t_1 \in (1,2)$; $t_2 \in (2,3)$ are implied, represented by the light-shaded rectangle. For Patient 2, the disease is detected after 2 years in the right eye, but is undetected in the left eye. Hence the implied intervals for Patient 2 are $t_1 \in (3,\infty)$; $t_2 \in (1,2)$, represented by the dark-shaded rectangle.

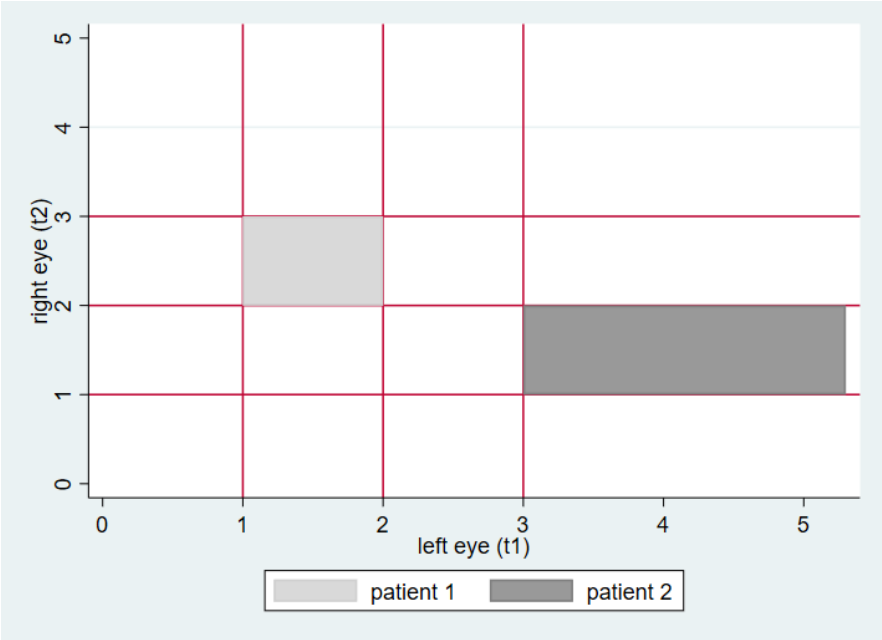

Figure 4.3: Regions implied for timing of onset of disease in left eye (t_1) and right eye (t_2) for two hypothetical patients, in the framework of Sun and Ding (2019).

The important point about the example illustrated in Figure 4.3 is that the regions of $(t_1\ ,\ t_2)$ space implied by the data from a single patient are rectangles with horizontal and vertical sides. Hence, given assumptions about the bivariate distribution of t_1 and t_2, it is a relatively simple matter to obtain the probability of falling in the implied region, this probability being the likelihood contribution used in the estimation of the parameters of the bivariate

[29] Probability weighting is one of the two concepts that are central to Prospect Theory (Kahneman and Tversky, 1992), the most popular framework for modelling departures from EU. The other is loss aversion. For a more detailed discussion of the concept of probability weighting, see Section 6.2.

distribution. This point is important because for the application of interest in this section, the bivariate data implies regions that are non-rectangular with curved sides. This gives rise to a rather interesting econometric challenge.

The problem of the simultaneous estimation of risk attitude and probability weighting using two MPL series has previously been considered by a number of researchers including Tanaka et al. (2010). In those studies, approximate values for the two parameters are obtained by taking the midpoints of the ranges applying to the two parameters. While computationally straightforward, this approach has a number of shortcomings. First, the method for obtaining the two parameter values can only be seen as a rough approximation, inducing the problem of measurement error. Second, observations corresponding to switches occurring at the beginning or the end of the MPL give rise to "open intervals", for which the choice of mid-point is arbitrary. This choice can have an important effect on estimation if a sizeable proportion of the observations fall in these open intervals. Finally, and most importantly, the process of replacing intervals with mid-points amounts to the discarding of a significant amount of information, and this inevitably results in a loss of precision in the estimation of the parameters.

Conte et al. (2019) use data from two MPL's to estimate the joint distribution of risk attitude and probability weighting in a way that makes maximal use of the available information. Their model is known as the Bivariate Random Preference (BRP) Model. The two MPL's are presented in the Appendix.

They assume the CRRA utility function $U(x) = x^{\alpha}$, $\alpha>0$, for which a value of $\alpha<1$ represents risk aversion, $\alpha=1$ represents risk neutrality, and $\alpha>1$ represents risk seeking. They assume the probability weighting function due to Prelec (1998), $w(p)=\exp(-(\ln(p))^{\gamma})$. If $\gamma=1$, we have EU. If $\gamma<1$, the weighting function is inverse S-shaped, with low-probability outcomes being overweighted, and high probability outcomes underweighted.[30] Since both parameters are strictly positive, a bivariate lognormal distribution (between-subject) is assumed:

$$\begin{pmatrix} \ln(\alpha_i) \\ \ln(\gamma_i) \end{pmatrix} = \mathrm{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\gamma \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\gamma \\ \rho\sigma_\alpha\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \right) \tag{4.14}$$

Conte et al.'s (2019) objective is to use the switchpoint data from the two MPLs to estimate the distributional parameters in (4.14), with the mean parameters μ_α and μ_γ being allowed to depend on subject characteristics and treatment variables.

For each of the two MPL's (see Appendix), the subject reports a switch point. The combination of two switch points indicates a set of combinations of the two parameters α and γ. These sets are shown in Figure 4.4 (left panel). It is important to note that, as alluded to earlier, the boundaries of the sets shown in Figure 4.4 are not, as in Figure 4.3, horizontal or vertical and nor are they even straight lines. The region may be described as a curvilinear quadrilateral. It is this feature that gives rise to the aforementioned modelling challenge. To estimate the parameters of (4.14) consistently, a likelihood function needs to be constructed in which each contribution is the volume under the bivariate lognormal density function over the irregularly-shaped region indicated by the observed pair of switch points. Conte et al. (2019) use Monte Carlo integration (Hammersley and Handscomb, 1964) for this purpose.

[30] See Section 6.2.for a more detailed discussion of the concept of probability weighting.

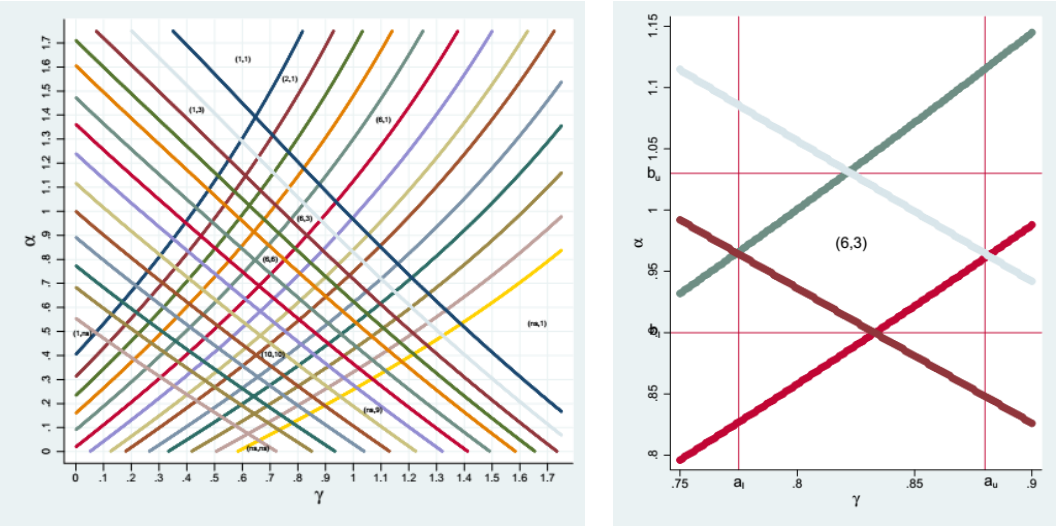

Figure 4.4. Left panel: Areas indicated by pairs of switch-points in the design of Conte et al. (2019). The numbers in brackets indicate the switch points in the first and second sequence. *ns* indicates the "no switch" cases. Right panel: Magnification of the area identified by the switch-point combination (6, 3).

A problem encountered when using a straightforward Monte Carlo integration technique for this purpose is rare-event sampling cases (see, for example, Rubinstein and Kroese, 2007). Some of the regions shown in Figure 4.4 (left panel) may be too small for the required probability to be estimated with any accuracy, with a modest number of simulations. For this reason, the importance sampling technique (Gourieroux and Monfort, 1996) is used. To illustrate, consider the case displayed in the right panel of Figure 4.4, which magnifies one of the curvilinear quadrilaterals from the left panel. A rectangle is drawn in the figure which is the smallest rectangle that completely contains the curvilinear quadrilateral. Under importance sampling, simulation is restricted to this rectangle in the evaluation of the required volume, leading to huge computational savings.

Conte et al. (2019) find that this consistent estimation method gives markedly different results from those obtained when OLS regression on midpoints is used.

One possible drawback with the bivariate random preference model is that there is no allowance for subjects to make random errors in reporting their switch-points. For a given subject, the two switch-points are used to infer a single region in which the subject's parameter pair is assumed to lie with certainty. If a third MPL were added to the sequence, it would result in an over-identification problem in the sense that a subject's third switch-point could easily contradict the information provided by the first two. This sort of contradiction would need to be interpreted in terms of within-subject randomness in decision making, and a different sort of model would be required.

4.6 Continuous (exact) data

An obvious way to eliciting the risk attitude of a subject is to present them with a lottery, and ask them for their "certainty equivalent", that is, the amount of money such that they would be exactly indifferent between receiving this sum of money and playing the lottery (WTA), or, alternatively, the amount they would be willing to pay to play the lottery (WTP).

For example, if the lottery is:

$$(0.3, \$3.85;\ 0.7, \$0.10)$$

and the subject claims that their certainty equivalent is \$0.75, then we can deduce that their coefficient of relative risk aversion is exactly 0.41. This is because 0.41 is the only value of r that satisfies the following inequality:

$$0.3\frac{3.85^{1-r}}{1-r}+0.7\frac{0.10^{1-r}}{1-r}=\frac{0.75^{1-r}}{1-r} \tag{4.15}$$

Suppose we have the values of *r* elicited in this way for a sample of subjects. This is "exact" data in the sense that the value of *r* is exactly observed.

We return to the assumption:

$$r \sim N\left(\mu,\sigma^2\right) \tag{4.16}$$

How do we estimate μ and σ when exact data is available? The answer is very simple: we find the sample mean and sample standard deviation of the observed r. If we wish to allow r to depend on subject characteristics, so that, for example, (4.16) becomes:

$$r \sim N\left(\beta_0+\beta_1 age+\beta_2 male,\sigma^2\right) \tag{4.17}$$

then the parameters of (4.17) could be estimated directly by applying a least squares regression of r on age and male.

Apart from being easier to work with, exact data leads to greater estimation efficiency relative to discrete data (such as binary, interval or ordinal data). Knowledge of the exact value taken by a variable clearly embodies more information that the knowledge that the value lies in an interval.

However, there are a number of problems with the use of exact data. Firstly, it is well known that 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.[31] Hence there is likely to be an underestimation of the degree of risk aversion when this form of elicitation is used. Secondly, as is well known (see e.g. Isoni et al., 2011), the value that is reported depends on whether the framing is in terms of willingness-to-pay (WTP) or willingness-to-accept (WTA). Thirdly, it is hard to incentivise the elicitation of valuations. The most popular technique for this is the Becker-DeGroot-Marschak mechanism (Becker et al. 1964),[32] but this is often considered to be too complicated for subjects to understand, and to divert attention from the valuation task itself (Bardsley et al., 2010, Section 6.5).

[31] The tendency for subjects to provide valuations close to the expected value of the lottery is itself an explanation for the preference reversal phenomenon (Grether and Plott, 1979) considered earlier in section 2.13.

[32] The Becker-DeGroot-Marschak (BDM) scheme is described as follows. The subject is asked to place a value on the lottery (i.e. to report their certainty equivalent). They are told that after they have done this a random "price" will be generated. If the randomly generated price is higher than their reported valuation, they will be given an amount of money equal to this price, and they will not play the lottery; if the price is lower than their valuation, they will play the lottery.

This is clearly another case in which the idiosyncrasies of human decision making are pulling in the opposite direction from issues of statistical efficiency in informing the most suitable choice of experimental design.

4.7 Censored data

Censored data is a mixture of continuous and discrete data (see Maddala, 1983). It often arises in settings in which subjects are required to make a transfer (e.g. dictator games) or contributions to a public fund (e.g. public goods games). In these settings, it is obvious that there is a lower limit to the subject's contribution, usually zero,[33] and an upper limit which is the subject's endowment. The model that is required in this case is the 2-limit tobit model (Nelson, 1976). The importance of allowing for censoring is that when OLS is used in the presence of censored data, the slope estimates tend to be seriously biased towards zero.

Another model which has become popular in Experimental Economics is the hurdle model (see Moffatt, 2015, Chapter 11), mentioned in Section 2.12 above in the context of dictator game giving. This is a generalisation of the tobit model that allows two distinct types of zero observation: censored zeros, as in the tobit model; and "zero types". A zero type is a subject who is destined to contribute zero whatever the circumstances. In the context of public goods games, a zero type is labelled a "free rider", and in the context of dictator games, they are "selfish".

Engel and Moffatt (2012) use a hurdle model to test for the presence of a "house money effect" in the context of a public goods game. The effect of house money is seen to be significant in the first hurdle: specifically, house money makes a subject less likely to be a free rider. Hence the effect of house money is more than just an effect on behaviour; it has the effect of changing a subject from one type to another. This result is potentially important in the external validity debate, since it suggests that "house money" – nearly always present in economic experiments – causes subjects to adopt a different persona during experiments from that used in everyday decisions.

5 Structural Estimation of social preference parameters

In the context of Experimetrics, structural estimation is often taken to mean the use of experimental data to estimate the parameters of a representative subject's utility function. One particular type of utility function that has been the focus of a good deal of attention is the utility function over own payoff and other's payoff in the context of a social preference experiment. Clearly an individual's utility depends positively on their own payoff. And if they are altruistic, it also depends positively on the other player's payoff. The utility function conveys the subject's relative concern for self and other (that is, their degree of altruism), and also their concern for efficiency (that is, the importance they attach to total welfare of both players). The purpose of this section is to describe methods that have used to estimate the parameters of utility functions of this form.

5.1 The Modified Dictator Game

[33] An interesting recent development is the emergence of "take games" – dictator games in which some treatments allow dictators to take money away from the recipient, i.e. to "give" less than zero. See Bardsley (2008) and List (2007). Of course, the issue of censoring is still relevant, but the censoring point is no longer zero.

In the modified[34] dictator game of Andreoni and Miller (2002), each subject is given an endowment (m) which they are required to allocate between "self" and "other", with both of these "goods" having a "price". For example, if the price of giving to "other" is 0.5, the amount actually received by "other" will be twice the amount allocated.

The following variables are defined:

x_1 = amount received by self
x_2 = amount received by other
m = endowment
p_1 = price of x_1
p_2 = price of x_2

Although x_1 and x_2 are the two arguments of the dictator's utility function, $U(x_1,x_2)$ say, they are not the decision variables. The decision variables are:

p_1x_1 = amount directed to self
p_2x_2 = amount directed to other

The two decision variables are not both free variables. They are constrained by the budget constraint:

$$p_1x_1 + p_2x_2 \leq m \tag{5.1}$$

In Andreoni and Miller's (2002) experiment, there were 176 subjects, each of whom faced a sequence of decision problems in the form of budgets. Each budget had a different combination of m, p_1 and p_2. The sequence of budgets are presented in Table 5.1.

Budget	m	p_1	p_2	observations	Mean amount directed to other
1	40	0.33	1	176	8.02
2	40	1	0.33	176	12.81
3	60	0.5	1	176	12.67
4	60	1	0.5	176	19.40
5	75	0.5	1	176	15.51
6	75	1	0.5	176	22.68
7	60	1	1	176	14.55
8	100	1	1	176	23.03
9	80	1	1	34	13.5
10	40	0.25	1	34	3.41
11	40	1	0.25	34	14.76

Table 5.1: Andreoni and Miller's (2002) design

The task required for each budget is to decide how much of the endowment (m) to keep for oneself (p_1x_1), and how much to direct to the other player (p_2x_2). The mean of p_2x_2 is shown in the final column of Table 5.1. Note that budgets 7, 8 and 9 represent standard dictator games because both prices are unity. Note further that in these three tasks, the amount given by the dictator is around 20% of the endowment, in agreement with the average seen in the literature (Camerer, 2003).

[34] The term "modified dictator game" is typically used for dictator games in which the prices of giving and keeping take values other than one.

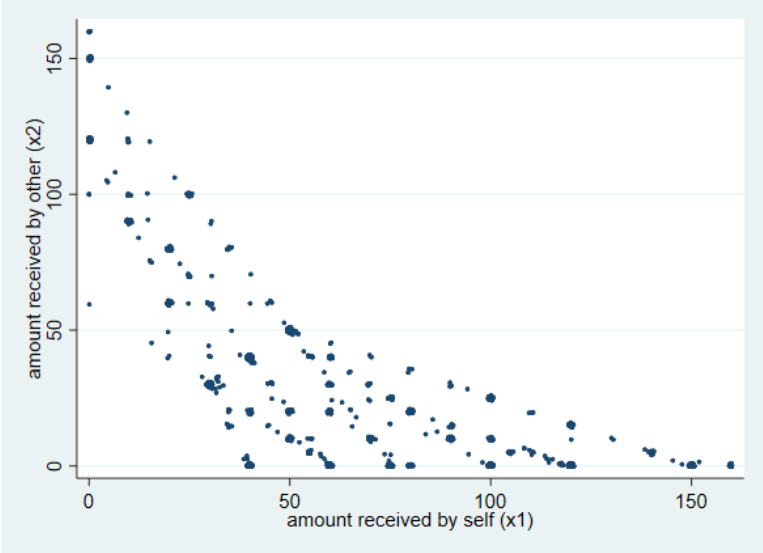

Figure 5.1: Jittered scatterplot of amount received by other (x_2) against amount received by self (x_1) in Andreoni and Miller's (2002) experiment.

Figure 5.1 shows a (jittered)[35] scatter plot of amount received by other against amount received by self from Andreoni and Miller's (2002) data. Note firstly that all points appear on one of 11 downward-sloping straight lines. These are, of course, the budget lines corresponding to the 11 budgets listed in Table 5.1. Note secondly that there is a concentration of points on the horizontal axis indicating selfishness, and a concentration of points on the 45-degree line ($x_2=x_1$) indicating concern for fairness The objective of the current exercise is to consider ways of using this data set to estimate the parameters of a utility function over own and other's payoff, and to interpret these parameters in terms of selfishness and concern for fairness.

5.2 Estimation of Social Preference Parameters

Following Andreoni and Miller (2002) and Jakiela (2013), we first assume the constant elasticity of substitution (CES) utility function:

$$U(x_1,x_2)=\left[\alpha x_1^{\rho}+(1-\alpha)x_2^{\rho}\right]^{\frac{1}{\rho}} \quad 0\le\alpha\le 1 \quad -\infty<\rho\le 1 \tag{5.2}$$

In (5.2), the parameter α indicates selfishness: it is the proportion of the budget that an individual would take for themselves in a standard dictator game. The parameter ρ indicates willingness to trade off equity and efficiency in response to price changes. It is more usual to interpret the elasticity of substitution, which is deduced from ρ using:

$$\sigma=\frac{1}{1-\rho} \tag{5.3}$$

Elasticity of substitution, σ, defined in (5.3), is between 0 and ∞. Values close to zero indicate that indifference curves are close to being L-shaped, and hence that all that matters to the individual is the equality of payoffs, that is, there is a concern for equity.

[35] A "jittered" scatterplot is a scatterplot in which the position of each point is randomly perturbed by a small amount. This is useful in situations in which multiple points occupy exactly the same position in the plot.

Values close to infinity indicate that indifference curves are downward sloping straight lines, and hence that all that matters to the individual is total payoff, that is, there is a concern for efficiency.

Maximising (5.2) subject to the budget constraint (5.1) leads to the Marshallian demand function for own payoff:

$$w_1 = \frac{p_1^{\frac{\rho}{1-\rho}}}{p_1^{\frac{\rho}{1-\rho}} + \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}} p_2^{\frac{\rho}{1-\rho}}} + \varepsilon \tag{5.4}$$

Where w_1 is the share of the total allocation that is allocated to "self", that is, $w_1 = \frac{p_1 x_1}{m}$. A stochastic term has been appended to (5.4) in order to turn the deterministic budget share equation into an estimable model. An equation for other's payoff is not needed because $w_1 + w_2 \equiv 1$.

It is possible to obtain estimates of the two parameters α and ρ by applying non-linear least squares to (5.4). That is:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\rho} \end{pmatrix}_{NLLS} = \underset{\alpha,\rho}{\arg\min} \sum_{i=1}^{n} \left(w_{1i} - \frac{p_{1i}^{\frac{\rho}{1-\rho}}}{p_{1i}^{\frac{\rho}{1-\rho}} + \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}} p_{2i}^{\frac{\rho}{1-\rho}}} \right)^2 \tag{5.5}$$

When applied to Andreoni and Miller's (2002) data, the estimates (and cluster-robust standard errors) thus obtained are $\hat{\alpha} = 0.69(0.02)$; $\hat{\rho} = 0.27(0.05)$. Applying (5.3) using the delta-method, we obtain an estimate of the elasticity of substitution $\hat{\sigma} = 1.37(0.09)$.

The estimate of α is significantly greater than 0.5, indicating a degree of selfishness. The estimate of the elasticity of substitution is significantly greater than 1, indicating that the subjects in this sample attach more importance to efficiency than to equity.

One feature of the data that is not incorporated in the above estimation is the accumulation of observations at the lower and upper limits of giving. Of the 1510 observations in Andreoni and Miller's (2002) data, 238(41.6%) represent zero giving, while 51 (3.4%) represent maximal giving. Hence the data is both upper and lower censored. Andreoni and Miller deal with this problem by estimating a 2-limit Tobit model (Nelson, 1976).

Moffatt and Zevallos-Porles (2020) go a step further by treating the boundary observations as corner solutions in the dictator's constrained optimisation problem, using a version of the Kuhn-Tucker theorem (Arrow and Enthoven, 1961). In their experiment, each dictator faces a sequence of tasks, in half of which only giving is possible, and in the remaining half only taking (from the recipient's endowment) is possible. A significant proportion of the observations lie at the extremes of the opportunity space, for example giving zero, or maximal taking, and this is why the Kuhn-Tucker approach is particularly apt. They focus on the difference between the estimated parameters in the giving and taking treatment. They find that the parameter representing selfishness is lower under the taking treatment, and this is interpreted in terms of the dictator being prepared to settle for a lower payoff when the task is a taking task, to allay the guilt associated with the task. In terms coined by Andreoni (1995), they are finding evidence that the "cold prickle" of taking is stronger than the "warm glow" of giving.

5.3 Estimation of Social Preference Parameters using Stated Choice Data

A very different approach to the estimation of social preference parameters is taken by Engelmann and Strobel (2004). They conduct surveys in which each respondent is asked to choose between three different (hypothetical) allocations, of the type shown in Table 5.2. Importantly, the respondent is given the identity of "Person 2".

Allocation	***A***	***B***	***C***
Person 1	8	6	10
Person 2	8	6	7
Person 3	4	6	7
Total	**20**	**18**	**24**

Table 5.2: An example of three hypothetical allocations similar to those used by Engelman and Strobel (2004). The survey respondent is given the identity of "Person 2".

The allocations in the example shown in Table 5.2 provide a useful example because all three have both attractions and drawbacks, and it is reasonable to expect the population to divide between the three when asked to choose between them. A purely selfish individual would choose allocation A, since, given their assumed identity as "Person 2", allocation A gives them the highest payoff. An individual who is strictly egalitarian is likely to choose B, since this allocation is the only one that gives rise to perfect equity. An individual who is concerned with efficiency is likely to choose C, since this is the allocation that gives the highest total payoff.

Engelmann and Strobel (2004) use the choices of 586 respondents to estimate a conditional logit model (Maddala, 1983). This model directly estimates the parameters of the utility function of a representative individual. The utility function assumed is (for allocation j):

$$U_j = \alpha_1 EFF_j + \alpha_2 MM_j + self_j + \alpha_3 FS\alpha_j + \alpha_1 FS\beta_j + \alpha_1 ERC_j + \varepsilon_j \equiv V_j + \varepsilon_j \qquad (5.6)$$

Where:

$$EFF_j = \sum_{k=1}^{3} x_{jk}$$

$$MM_j = \min\left\{x_{jk}, k = 1,2,3\right\}$$

$$SELF_j = x_{j2}$$

$$FS\alpha_j = -\tfrac{1}{2}\sum_{k\neq 2} \max\left\{x_{jk} - x_{j2}, 0\right\}$$

$$FS\beta_j = -\tfrac{1}{2}\sum_{k\neq 2} \max\left\{x_{j2} - x_{jk}, 0\right\}$$

$$ERC_j = -100\left|\frac{1}{3} - \frac{x_{j2}}{EFF_j}\right|$$

In (5.6), EFF is the measure of efficiency mentioned above. MM stands for minimax, and represents concern for the worst-off person. SELF is the individual's own payoff. FSα and FSβ are the two measures of inequity aversion distinguished famously by Fehr and Schmidt (1999). FSα represents concern for disadvantageous inequity (e.g. allocation C in Table 5.2), while FSβ represents concern for advantageous inequity (e.g. allocation A). Finally, ERC embodies the "Equity, Reciprocity and Competition" theory of Bolton and Ockenfels (2000). This measure penalises allocations for which the individual's own payoff is far (in either direction) from one third of the total.

In (5.6), V_j is the deterministic component of utility, and ε_j is the random component, and it is assumed that the alternative with the highest utility $U_j = \max(U_1, U_2, U_3)$ is chosen. For convenience, it is assumed that the ε_j's are independent and identically distributed with type I extreme value distribution (also known as the Gumbel distribution), defined by the density function:

$$f(\varepsilon) = \exp\left[-\varepsilon - \exp(-\varepsilon)\right] \quad -\infty < \varepsilon < \infty \tag{5.7}$$

With assumption (5.7), it is well known (see Maddala, 1983) that the probability of alternative j being chosen is:

$$P(y_j = 1) = \frac{\exp(V_j)}{\sum_{k=1}^{3} \exp(V_k)} \tag{5.8}$$

The likelihood function for the conditional logit model is constructed using (5.8).

The key findings of Engelman and Strobel (2004) are that EFF and MM are the most important determinants of utility, and they both have the expected sign in the utility function. On the basis of this hypothetical choice data set, they find little support for the theories of either Fehr and Schmidt (2002) or Bolton and Ockenfels (2000).

6. Continuous Heterogeneity: Maximum Simulated Likelihood

We will now return to the setting of choice under risk. However, while in earlier sections we were concerned with data sets containing only one decision per subject, now we are interested in the more usual setting in which each subject (i=1,…,n) faces a sequence of (t=1,…,T) choice problems. Econometrically, this leads us into the realms of panel data analysis.

The focus of this section will be methods for estimating the population distribution of risk attitude and also the probability weighting parameters, and the various stochastic specifications that are available for pursuing this objective.

6.1 Theoretical background for choice under risk

We commence with a brief outline of the theoretical concepts that are central to the modelling of choice under risk. A key feature of many such models is the utility function that is assumed. We first assume that the outcome of the task is an amount of income, or "wealth", which we shall label *x*. The utility function, *U(x)*, is defined over this variable. A number of different specifications of *U(x)* are possible. An important point is that the type of utility function required here is unique up to positive affine transformations. It is not permitted to apply any order-preserving transformation, as is possible for utility functions defined over consumption quantities. This essentially means that the *shape* of the utility function (e.g. its curvature) is important.

Two concepts, attributable to Pratt (1964), are central to the modelling of behaviour under risk: absolute risk aversion; and relative risk aversion. Both of these measures of risk aversion relate closely to the curvature of the utility function. The coefficient of absolute risk aversion is defined by:

$$A(x) = -\frac{U''(x)}{U'(x)} \tag{6.1}$$

The coefficient of relative risk aversion is defined by:

$$R(x) = -\frac{xU''(x)}{U'(x)} \tag{6.2}$$

Note that if the Utility function is linear, e.g. $U(x) = x$, then both of the above measures of risk aversion are zero, and the decision-maker is classified as risk neutral. We would normally expect risk aversion, that is, we expect both measures to take positive values. A risk-seeking decision-maker would have negative values for the two measures.

A utility function that is very popular in decision theory and also throughout the subject of economics is the constant relative risk aversion (CRRA) utility function. The most popular way of parameterising this utility function is as follows:

$$\begin{aligned} U(x) &= \frac{x^{1-\alpha}}{1-\alpha} \qquad x \geq 0;\ -\infty < \alpha < \infty;\ \alpha \neq 1 \\ &= \ln(x) \qquad \alpha = 1 \end{aligned} \tag{6.3}$$

Applying (6.2) to (6.3), it is seen that the coefficient of relative risk aversion is α. Note that the function appearing in the first line of (6.3) is not defined when α=1. The limit of the function as α approaches 1 is ln(x), and hence this is what is used when α=1.

A simpler way of parameterising CRRA is using the "Power" utility function, defined by:

$$\begin{aligned} U(x) &= x^r \qquad x \geq 0;\ r > 0 \\ &= \ln(x) \quad r = 0 \\ &= -x^r \qquad r < 0 \end{aligned} \tag{6.4}$$

Applying (6.2) to (6.4), it is seen that the coefficient of relative risk aversion under this parameterisation is 1-r.

One problem with both of the CRRA functions (6.3) and (6.4) is that they do not fully accommodate zero outcomes. It is very common for the lowest outcome in a lottery to be zero, and hence we need to be able to evaluate U(0). However, when the power parameter is non-positive, i.e. if $\alpha \geq 1$ in (6.3) or r<0 in (6.4), U(0) is not defined. This does not mean that the CRRA utility function cannot be used when the lowest outcome is zero. It simply means that, when the lowest outcome is zero, the function is incapable for explaining high degrees of risk aversion.

A different sort of utility function that avoids this problem is the Constant Absolute Risk Aversion (CARA) function, defined by:

$$U(x) = 1 - \exp(-rx) \qquad x \geq 0;\ r > 0 \tag{6.5}$$

Applying (6.1) to (6.5), it is seen that the coefficient of absolute risk aversion r. Note that in (6.5), r must be positive to ensure that U(x) is increasing in x. Note also that (6.5) is defined for all values of x including zero. Hence the CARA function avoids the problem identified above relating to the CRRA: the CARA is able to explain high degrees of risk aversion even when the lowest outcome is zero.

A more general utility function that has become popular in applications is the expo-power utility function (Saha, 1993; Holt and Laury, 2002), defined as follows:

$$U(x) = 1 - exp(-\beta x^{\alpha}) \quad x \geq 0;\ \alpha \neq 0;\ \beta \neq 0;\ \alpha\beta > 0 \qquad (6.6)$$

The attractive feature of (6.6) is that is combines CARA and CRRA, with absolute risk aversion represented by α, and relative risk aversion by β. However, note that $\beta \neq 0$, which implies that exact CRRA is not included as a special case.

6.2 Decision-Theoretical Framework: EU and RDU

Here we introduce a general framework and a set of notational conventions that are useful for the modelling of choice under risk. Comparable frameworks have been used by Hey and Orme (1994), Loomes et al. (2002), Conte et al. (2011), and others.

Let us assume that all choice tasks in the experiment involve combinations of the M money outcomes: $0 \leq x_1 < x_2 < \ldots < x_M$. We index the tasks by *t* (*t*=1,…,*T*). For most choice tasks, one of the two lotteries may be classified as the "riskier" lottery, and the other as the "safer" lottery. When this is not possible, the task is one of "dominance", since one lottery first-order stochastically dominates the other.[36] If task *t* is a non-dominance task, we will label the riskier lottery as $\mathbf{p_t}$, and the safer as $\mathbf{q_t}$. For a dominance task, $\mathbf{p_t}$ will be the dominating lottery, and $\mathbf{q_t}$ the dominated. $\mathbf{p_t} = (p_{1t}\ p_{2t}\ \ldots p_{Mt})'$ and $\mathbf{q_t} = (q_{1t}\ q_{2t}\ \ldots q_{Mt})'$ are vectors containing probabilities corresponding to the M possible outcomes.

We index subjects by *i* (*i*=1,…,*n*), and, for the sake of convenience, we shall assume the power utility function introduced in (6.4) above, for subject i:

$$U_i(x) = x^{r_i} \qquad x \geq 0;\ r_i > 0 \tag{6.7}$$

Recall that (6.7) is a simple version of the Constant Relative Risk Aversion (CRRA) utility function, and the coefficient of relative risk aversion (for subject *i*) is $1 - r_i$.

Under the assumption of Expected Utility (EU) maximisation, subject *i*'s valuations of the two lotteries in choice problem *t* are, using (6.7):

[36] To establish dominance, it is necessary to obtain, for each lottery, the function representing the probability of receiving an outcome of at least X as a function of X. One lottery dominates the other if its function so-obtained is sometimes above, and never below, that of the other lottery.

$$V_i^{EU}(\mathbf{p_t};r_i) \equiv \sum_{m=1}^{M} p_{mt} U_i(x_m) = \sum_{m=1}^{M} p_{mt} x_m^{r_i}$$
$$V_i^{EU}(\mathbf{q_t};r_i) \equiv \sum_{m=1}^{M} q_{mt} U_i(x_m) = \sum_{m=1}^{M} q_{mt} x_m^{r_i} \quad (6.8)$$

We define the *valuation differential* of $\mathbf{p_t}$ over $\mathbf{q_t}$ under EU, $\nabla_{it}^{EU}(r_i)$, as follows.

$$\nabla_{it}^{EU}(r_i) \equiv V_i(\mathbf{p_t};r_i) - V_i(\mathbf{q_t};r_i) \quad (6.9)$$

Under a deterministic version of EU, subject *i* chooses $\mathbf{p_t}$ over $\mathbf{q_t}$ if this *valuation differential* is positive, that is, if $\nabla_{it}^{EU}(r_i) > 0$.

The decision theory literature contains a considerable volume of evidence of violations of EU (see Starmer, 2000). A natural way of relaxing EU within the present framework is to assume a form of Rank-Dependent Utility (RDU) theory (Quiggin, 1982). RDU Theory is based on the concept of a probability weighting function, which is one of the key building blocks of Prospect Theory (Kahneman and Tverky, 1979; Tversky and Kahneman, 1992).

As above, consider the lottery $\mathbf{p} = (p_1\ p_2\ \ldots p_M)'$. According to RDU, a decision maker does not use the true probabilities contained in **p** when evaluating the lottery, but rather uses the transformed probabilities $\tilde{p}_1, \tilde{p}_2, \cdots, \tilde{p}_M$. The transformed probabilities are obtained from the true probabilities as follows:

$$\tilde{p}_j = w\left(\sum_{k=j}^{M} p_k\right) - w\left(\sum_{k=j+1}^{M} p_k\right) \quad j = 1 \cdots M-1$$
$$\tilde{p}_M = w(p_M) \quad (6.10)$$

where w(.) is the weighting function, with the properties w(0)=0, w(1)=1. Note that (6.10) ensures that the transformed probabilities sum to one, i.e. that $\sum_{j=1}^{M} \tilde{p}_j = 1$.

It is useful to consider the case with only three possible outcomes, for which (6.10) becomes:

$$\tilde{p}_3 = w(p_3)$$
$$\tilde{p}_2 = w(p_2 + p_3) - w(p_3) \quad (6.11)$$
$$\tilde{p}_1 = 1 - w(p_2 + p_3)$$

The three-outcome case is illustrated in Figure 6.1, in which the curve is the weighting function. The inverse-S shape of this curve causes the probabilities of the best and worst outcomes to be overweighted (i.e. $\tilde{p}_3 > p_3$ and $\tilde{p}_1 > p_1$ respectively), implying that the probability of the middle outcome is underweighted (i.e. $\tilde{p}_2 < p_2$). Note that, if the weighting function coincided with the 45^0-line (i.e. if w(p)=p), the probabilities would be correctly weighted, and we would be back to EU.

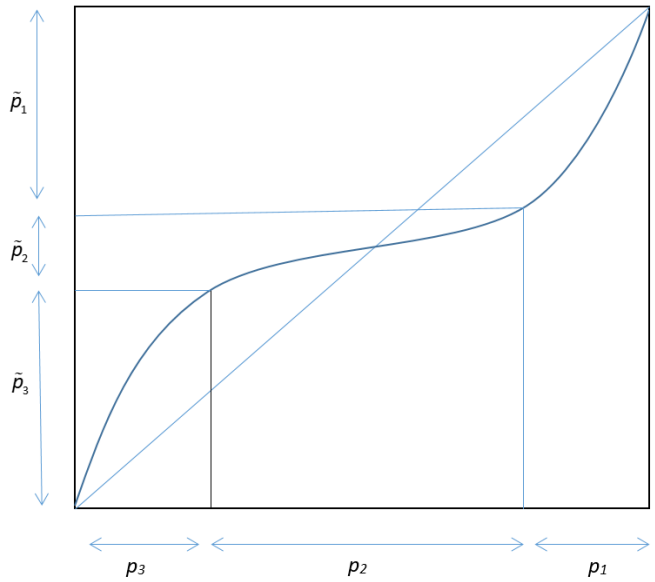

Figure 6.1: An inverted S-shaped weighting function. Three outcomes assumed.

Three popular parametric weighting functions are specified as follows:[37]

Power: $w(p) = p^{\gamma}$, with $\gamma > 0$

Tversky and Kahneman (1992): $w(p) = \dfrac{p^{\gamma}}{\left(p^{\gamma} + (1-p)^{\gamma}\right)^{1/\gamma}}$, with $\gamma > 0.279$

Prelec (1998): $w(p) = \exp\left(-\alpha\left(-\ln p\right)^{\gamma}\right)$, with $\alpha > 0; \gamma > 0$

The first of these, the power weighting function, might be seen as undesirably restrictive since it does not allow an inverse S-shape; it is either completely above (if γ<1) or completely below (if γ>1) the 45° line. The second function is due to Tversky and Kahneman (1992). Despite having only one parameter, this function has the required inverse-S shape (if 0.279<γ<1). The lower limit of γ is required for monotonicity. We have EU when γ = 1. The third function is due to Prelec (1998) and also has an inverse-S shape. It has two parameters.[38] The parameter α reflects pessimism (with values less than 1 indicating optimism) and the parameter γ determines the pronouncedness of the inverse-S shape (the lower the value of γ, the more pronounced the inverse-S; γ>1 would instead indicate an S-shape). When both of these parameters are equal to one, we have EU. It is typically found that both parameters are somewhat less than one.

It is reasonable to assume that probability weighting, like risk aversion, varies between subjects. For simplicity, let us assume there is one weighting parameter γ which is γ_i for subject i. Having used the chosen weighting function to find the transformed probabilities of the two lotteries $\mathbf{p}_t$ and $\mathbf{q}_t$ for subject i, we may compute this subject's valuation of each (under RDU) as:

[37] A useful survey of probability weighting functions appearing in the literature is provided by Stott (2006).

[38] α is sometimes referred to as the "elevation parameter" and is said to reflect "attractiveness" (Gonzalez and Wu 1999), while γ reflects "a-insensitivity" (Tversky and Fox 1995).

$$V_i^{RDU}\left(\mathbf{p_t};r_i,\gamma_i\right)\equiv\sum_{m=1}^{M}\tilde{p}_{mt,i}\left(\gamma_i\right)U_i\left(x_m\right)=\sum_{m=1}^{M}\tilde{p}_{mt,i}\left(\gamma_i\right)x_m^{r_i}$$
$$V_i^{RDU}\left(\mathbf{q_t};r_i,\gamma_i\right)\equiv\sum_{m=1}^{M}\tilde{q}_{mt,i}\left(\gamma_i\right)U_i\left(x_m\right)=\sum_{m=1}^{M}\tilde{q}_{mt,i}\left(\gamma_i\right)x_m^{r_i} \quad (6.12)$$

As with EU, we define the *valuation differential* of $\mathbf{p_t}$ over $\mathbf{q_t}$ under RDU, $\nabla_{it}^{RDU}\left(r_i,\gamma_i\right)$, as follows.

$$\nabla_{it}^{RDU}\left(r_i,\gamma_i\right)\equiv V_i\left(\mathbf{p_t};r_i,\gamma_i\right)-V_i\left(\mathbf{q_t};r_i,\gamma_i\right) \quad (6.13)$$

Under a deterministic version of RDU, subject *i* chooses $\mathbf{p_t}$ over $\mathbf{q_t}$ if $\nabla_{it}^{RDU}\left(r_i,\gamma_i\right)>0$.

To allow estimation, we need to incorporate a stochastic element to (6.9) or (6.13). Broadly, there are two approaches to this: the Fechner approach, and the random preference (RP) approach. Here, we focus on the Fechner approach; the RP approach is considered later in Section (6.10).

6.3 The Fechner Model (Random Utility Model)

Consider the RDU model developed in the last sub-section. Assume that each subject (i) has a risk-attitude parameter (r_i) and probability weighting parameter(s) (γ_i say) that are constant between tasks. The Fechner model (Fechner, 1860; Hey and Orme, 1994), also known as the Random Utility Model, is obtained by appending a homoscedastic error term to (6.13), whereupon the condition for the choice of $\mathbf{p_t}$ becomes:

$$\nabla_{it}^{RDU}\left(r_i,\gamma_i\right)+\varepsilon_{it}>0$$
$$\varepsilon_{it}\sim N\left(0,\sigma^2\right). \quad (6.14)$$

Defining the binary indicator y_{it} =1(-1) if subject i chooses $\mathbf{p_t}$ ($\mathbf{q_t}$), the probabilities of the two choices (conditional on subject-specific parameters) are given by:

$$P\left(y_{it}\,|\,r_i,\gamma_i\right)=\Phi\left(y_{it}\times\frac{\nabla_{it}^{RDU}\left(r_i;\gamma_i\right)}{\sigma}\right) \quad (6.15)$$

where $\Phi(.)$ is the standard normal c.d.f.

6.4 The tremble parameter

The Fechner model clearly allows subjects to make "incorrect" choices when they are close to indifference (i.e. when the valuation differential is close to zero). However, assuming that the noise parameter, σ, is relatively small, the Fechner model has problems explaining major errors, such as dominance violations.[39] The choice of a dominated alternative may require an astronomically high value of the Fechner error term, and hence dominance violations appearing in the data have the potential to distort estimates. This leads us to consider the addition to the model of a different type of error: the "tremble".

[39] Loomes and Sugden (1998) report that around 1.5% of choices made in dominance problems are violations of dominance.

We shall follow Loomes et al (2002) by introducing a tremble parameter ω to the choice model. ω is the probability that on any task a subject loses concentration and chooses randomly between the two alternatives, with equal probability. Issues surrounding the estimation of and testing for tremble effects have been discussed in some detail by Moffatt and Peters (2001).

Applying this extension to the Fechner model (6.15), we obtain:

$$P\left(y_{it} \mid r_i, \gamma_i\right) = \left(1-\omega\right)\Phi\left(y_{it} \times \frac{\nabla_{it}^{RDU}\left(r_i;\gamma_i\right)}{\sigma}\right) + \frac{\omega}{2} \tag{6.16}$$

6.5 The role of experience

It is possible that subjects' behaviors may change systematically in the course of the experiment, revealing the effect of experience. In order to allow for this, certain parameters may be allowed to depend on the amount of experience accumulated, which may be measured using the position of the current task in the sequence of tasks. For this purpose we define a variable τ_{it} to be the position of task *t* in the sequence of tasks undertaken by subject *i*. Note that a sensible feature of a design is that each subject encounters the sequence of tasks in a different order, that is, $\tau_{it} \neq \tau_{jt}$, $i \neq j$. This feature is a generalisation to the multiple-task setting of the concept of the "crossover design" used in a two-task setting and discussed in Section 3.2 above.

One of Loomes et al.'s (2002) most striking findings is the change with experience in their probability weighting parameters. Their key weighting parameter (termed the "bottom-edge effect", and labeled b) is one that represents the over-weighting of a small probability of the best outcome. Under EU, b=0. They allow this parameter to depend on task number with the following specification:

$$b_{it} = \beta_0 \exp\left(\beta_1 \tau_{it}\right) \quad t = 1, \cdots, T \tag{6.17}$$

The parameter β_0 in (6.17) is interpreted as the probability weighting parameter at the start of the experiment, that is, when the subject has zero experience. The parameter β_1, if negative, represents the rate of decay towards zero of the parameter. Loomes et al. (2002) indeed find that β_1 is significantly negative, implying that probability weighting decays dramatically in the course of the experiment, although not all the way to zero. They conclude that decision makers use distorted probabilities in early tasks, but undergo a learning process that moves them closer to EU with experience.

Another parameter which it may be considered natural to allow to change with experience is the tremble parameter. This parameter has been defined as the probability that a subject loses concentration when undertaking a particular task. This interpretation may be broadened: another reason why a subject might choose randomly between the two alternatives is a lack of understanding of the task. We might hypothesise that such a lack of understanding is more likely to arise earlier in the sequence than later; hence we would expect the tremble probability to start off at a relatively high value, but to decay towards zero over the course of the experiment.

A suitable specification would therefore be:

$$\omega_{it} = \omega_0 \exp\left(\omega_1 \tau_{it}\right) \tag{6.18}$$

In (6.18), ω_0 represents the tremble probability at the start of the experiment, while ω_1 (assuming it is negative) represents the speed of its decay. The specification (6.18) has been usefully applied by Loomes et al. (2002) and Bardsley and Moffatt (2007).

6.6 Between-subject variation and the sample log-likelihood

The Fechner models under EU and RDU were constructed above in terms of the behavior of an individual subject. The choice probabilities were derived, conditional on the subject's preference parameters, r_i (and γ_i in the case of RDU). It is of course essential to allow for between-subject variation in one or both of these preference parameters.

Let us consider the Fechner-RDU model with tremble, defined in (6.16) above. Since both risk aversion (r_i) and probability weighting (γ_i) parameters are constrained to be positive, it is natural to assume that they vary over the population according to a bivariate lognormal distribution:

$$\begin{pmatrix} \ln r \\ \ln \gamma \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \eta_1^2 & \rho\eta_1\eta_2 \\ & \eta_2^2 \end{pmatrix} \right] \tag{6.19}$$

The likelihood contribution associated with the T decisions of subject i is:

$$L_i = \int_0^\infty \int_0^\infty \prod_{t=1}^{T} \left[\left(1-\omega_{it}\right) \Phi\left(y_{it} \times \frac{\nabla_{it}^{RDU}\left(r,\gamma\right)}{\sigma} \right) + \frac{\omega_{it}}{2} \right] f\left(r,\gamma;\mu_1,\mu_2,\eta_1,\eta_2,\rho\right) dr\, d\gamma \tag{6.20}$$

where $f(r, \gamma; \mu_1, \mu_2, \eta_1, \eta_2, \rho)$ is the bivariate lognormal density function associated with (6.19).

The sample log-likelihood function is then:

$$LogL = \sum_{i=1}^{n} L_i \tag{6.21}$$

The model defined in (6.19)-(6.21) is a version of the random effects probit model, of the type estimated by Conte et al. (2010, 2018).

6.7 The method of Maximum Simulated Likelihood (MSL)

The key practical problem encountered when attempting to maximise likelihood functions such as that defined in (6.16)-(6.18) is how to evaluate the double integral appearing in (6.17). It is a double integral because two dimensions of subject-heterogeneity are assumed: risk aversion and probability weighting. A method that has become very popular in recent years is based on simulation, and the resulting estimation method is known as the method of maximum simulated likelihood (MSL, see Train, 2003).

The method of MSL essentially involves replacing the integral with a mean of values of the function evaluated at a set of simulated values. The choice of simulation routine is commonly a Halton sequence (Halton, 1960).

Let I be the value of the double integral appearing in the log-likelihood for the Fechner RDU model (6.20):

$$I = \int_0^\infty \int_0^\infty \prod_{t=1}^{T} \left[(1-\omega_{it}) \Phi \left(y_{it} \times \frac{\nabla_{it}^{RDU}(r,\gamma)}{\sigma} \right) + \frac{\omega_{it}}{2} \right] f(r,\gamma;\mu_1,\mu_2,\eta_1,\eta_2,\rho) dr d\gamma \tag{6.22}$$

The method of evaluation of I in (6.22) is to use the following average as an approximation to it:

$$\hat{I} = \frac{1}{J} \sum_{j=1}^{J} \prod_{t=1}^{T} \left[(1-\omega_{it}) \Phi \left(y_{it} \times \frac{\nabla_{it}^{RDU}(r_j,\gamma_j)}{\sigma} \right) + \frac{\omega_{it}}{2} \right] \tag{6.23}$$

where r_j, $j=1,\ldots,J$; γ_j, j=1,…J, are two sets of Halton draws, transformed in such a way as to resemble variates from the bivariate lognormal distribution specified in (6.19). Note that the double integral in (6.22) has been replaced by a single summation in (6.23). Even if the dimensionality of the integral were greater than two, it would still be replaced by a single summation, and this is the sense in which the method of MSL overcomes the "curse of dimensionality".

Halton (1960) draws, in raw form, follow a uniform distribution. To highlight their properties it is informative to plot two Halton sequences against oneanother.[40] Such a plot is shown in the right panel of Figure 6.2. In the left panel is shown a pair of random uniforms, obtain using a pseudo-random number generator.

The two properties of Halton draws that make them particularly useful for numerical integration of the type being considered here are coverage and covariance. The coverage issue is as follows. With independent random draws (see left panel of Figure 6.2) it is inevitable that draws are clumped together in some areas, with a scarcity of draws in other areas. In contrast, the right panel of Figure 6.2 clearly illustrates the property of even coverage of Halton draws, and draws with even coverage are known to provide a more accurate approximation to the integral.

The covariance issue is as follows. With independent draws, the autocovariance between draws is zero. The variance of a simulator would therefore be the variance based on a single draw divided by J (the number of draws). Halton draws are negatively autocorrelated instead of independent, that is, a high draw tends to be followed by a low draw, and vice versa. This means that a draw that gives rise to an over-estimate of I will be immediately followed by one that gives an under-estimate, and vice versa, and this gives rise to a lower variance of the average, and hence faster convergence of the average to the true I.

The uniform draws need to be transformed in accordance with the distributional assumptions regarding the random terms within the model. If the two uniform sequences are u_1 and u_2, they would be transformed to r and γ, whose distribution are specified in (6.19), using:

$$\begin{aligned} r &= \exp\left(\mu_1 + \eta_1 \Phi^{-1}(u_1) \right) \\ \gamma &= \exp\left(\mu_2 + \eta_2 \left[\rho \Phi^{-1}(u_1) + \sqrt{1-\rho^2}\, \Phi^{-1}(u_2) \right] \right) \end{aligned} \tag{6.24}$$

[40] Here "two Halton sequences" means, for example, the Halton sequence used to represent (after being transformed) the risk aversion of subject 1, and that used to represent that of subject 2.

Figure 6.3 shows two series for r and γ generated using (6.24).

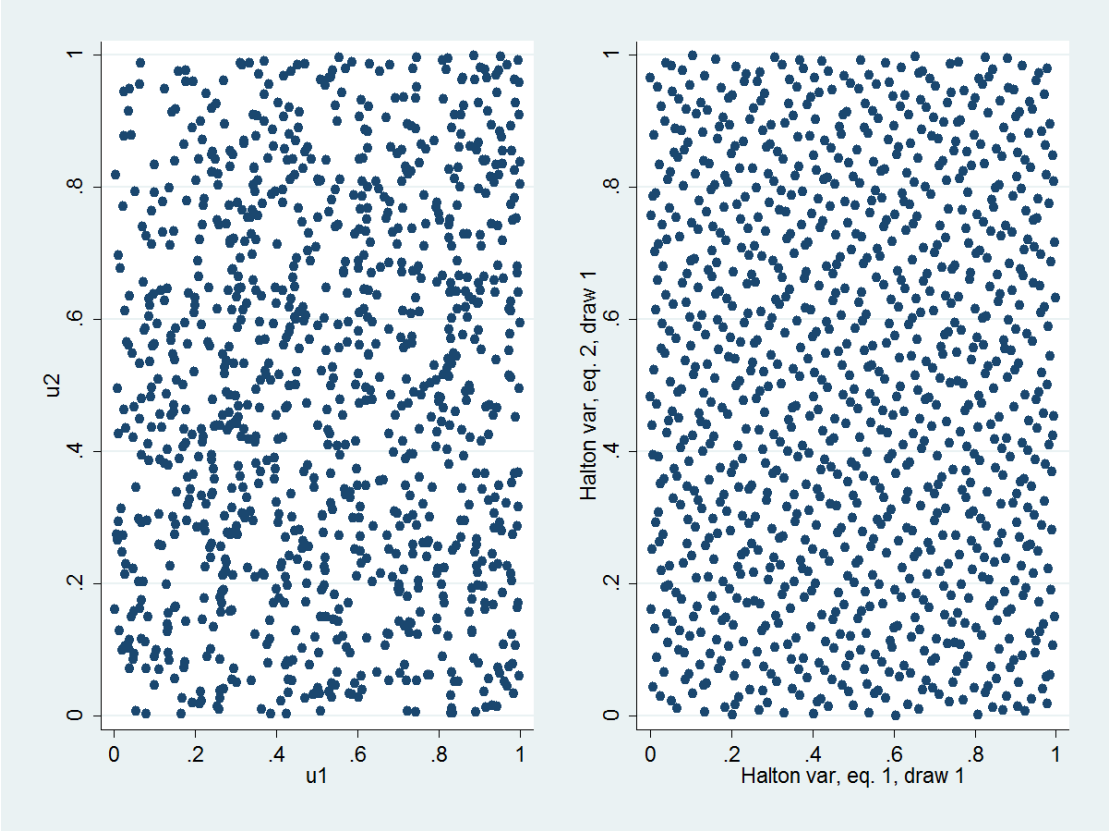

Figure 6.2. Left panel: two random uniforms; Right panel: two Halton sequences

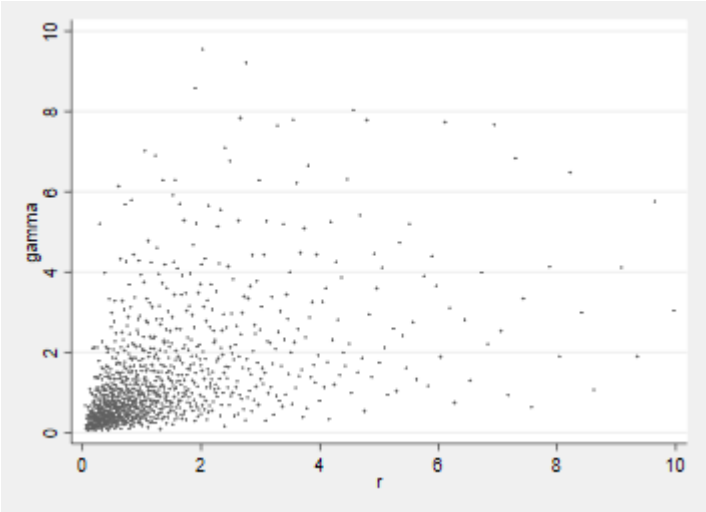

Figure 6.3. Two Halton sequences transformed to two log-normals with (underlying) correlation +0.60. Both means zero; both s.d.s 1.

Further detail on the use of MSL to estimate risky choice models, including complete STATA programs, is provided by Moffatt (2015). The sort of risky choice model considered here has been estimated using MSL by Conte et al. (2010, 2018). In both cases they find wide heterogeneity in both risk aversion and probability weighting, and in both cases they find a

positive correlation: more risk-averse subjects exhibit greater probability distortion. Von Gaudecker et al. (2011) go further and assume between-subject heterogeneity in four different preference parameters: utility curvature, loss aversion, preferences toward the timing of uncertainty resolution, and the propensity to choose randomly rather than on the basis of preferences. Estimation of such a model by MSL clearly requires four sets of Halton draws.

6.8 Post-Estimation

Having estimated the Fechner-RDU model, it is possible to obtain posterior estimates of the preference parameters for each subject in the sample. This is done using Bayes' Rule. For example, the posterior estimate of risk attitude for subject i is:

$$\hat{r}_i = \frac{\int_0^\infty \int_0^\infty r \prod_{t=1}^{T} \left[(1-\hat{\omega}_{it}) \Phi\left(y_{it} \times \frac{\hat{\nabla}_{it}^{RDU}(r,\gamma)}{\hat{\sigma}} \right) + \frac{\hat{\omega}_{it}}{2} \right] f(r,\gamma;\hat{\mu}_1,\hat{\mu}_2,\hat{\eta}_1,\hat{\eta}_2,\hat{\rho}) \, dr \, d\gamma}{\int_0^\infty \int_0^\infty \prod_{t=1}^{T} \left[(1-\hat{\omega}_{it}) \Phi\left(y_{it} \times \frac{\hat{\nabla}_{it}^{RDU}(r,\gamma)}{\hat{\sigma}} \right) + \frac{\hat{\omega}_{it}}{2} \right] f(r,\gamma;\hat{\mu}_1,\hat{\mu}_2,\hat{\eta}_1,\hat{\eta}_2,\hat{\rho}) \, dr \, d\gamma} \quad (6.25)$$

where hats indicate that parameters have been replaced by estimates.

6.9 The Random Preference (RP) Model

The Random Preference (RP) Model was introduced by Loomes and Sugden (1998) and subsequently estimated econometrically by Loomes et al. (2002) and others. It is an alternative approach to the Fechner approach outlined above, and it is partly motivated by a problem with the Fechner model has recently come to the fore in the decision theory literature (Wilcox, 2011; Blavatskyy, 2011; Apesteguia and Ballester, 2016).

This problem with the Fechner model is that when risk aversion reaches high values, the valuation differential (∇) approaches zero, and the choice probabilities approach one half. A consequence of this is that the probability of choosing the risky alternative may not be a monotonically decreasing function of risk aversion. This leads to a number of practical problems, most obviously, an identification problem arising from the fact that the same choice probabilities may be associated with two different levels of risk aversion.

A simple example is useful to illustrate this problem. Consider the choice between a 50:50 gamble of 1 and 3 (the risky lottery), and a certainty of 2 (the safe lottery). Assume that the decision maker has power utility, is an EU maximiser, and makes decisions subject to a Fechner error with variance 0.04. Their probability of choosing the risky lottery against their risk attitude (power) parameter is shown in Figure 6.4. Logically, we expect the probability of choosing the risky alternative to be monotonically increasing in the power parameter. As seen in Figure 6.4, this is definitely not the case, and hence the problems raised in the last paragraph are clearly seen even in the context of this very simple example.

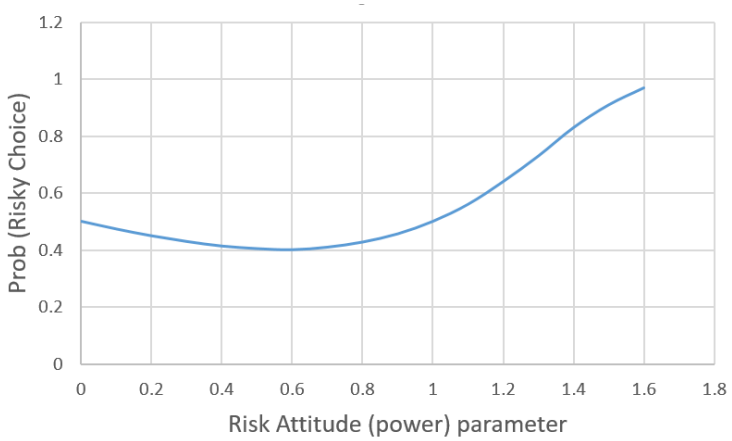

Figure 6.4: Probability of risky choice against risk attitude (power) parameter, assuming EU. Risky choice: (0.5, \$1; 0.5, \$3); Safe choice: (1, \$2). Fechner error variance = 0.04.

One way of viewing the Fechner model is to recognise that each of the two alternatives in the choice problem has an error term, and these two errors distort the utilities independently. An alternative modelling approach is to allow within-subject variation in the preference parameters, but to assume that for given values of these parameters, the choice problem is solved without error. This strategy avoids the problem identified in the previous paragraphs because the evaluation of each alternative is not distorted independently of the other.

This alternative modelling strategy is known as the Random Preference (RP) approach (Loomes and Sugden, 1998; Loomes et al., 2002), explained as follows. For the sake of simplicity, let us assume that subjects are EU maximisers, and that only a subject's risk-attitude parameter (r) varies between tasks. So, we assume that the risk attitude of subject i in task t comes from the distribution:

$$\ln(r_{it}) \sim N(m_i, \sigma^2) \tag{6.26}$$

The parameter σ^2 in (6.26) represents the within-subject variance in risk attitude. We are further assuming that each subject i has a different "mean" risk attitude, m_i.

Recall that the valuations of the two lotteries (assuming EU) are given by:

$$\begin{aligned} V_i^{EU}(\mathbf{p_t}; r_{it}) &\equiv \sum_{m=1}^{M} p_{mt} U_{it}(x_m) = \sum_{m=1}^{M} p_{mt} x_m^{r_{it}} \\ V_i^{EU}(\mathbf{q_t}; r_{it}) &\equiv \sum_{m=1}^{M} q_{mt} U_{it}(x_m) = \sum_{m=1}^{M} q_{mt} x_m^{r_{it}} \end{aligned} \tag{6.27}$$

In the context of the RP model, what is required in order to proceed with estimation is, for each task t, a range of values of r for which the riskier lottery $\mathbf{p_t}$ will be chosen over the safer lottery $\mathbf{q_t}$. That is, we require to find the value r_t* for which:

$$r_{it} > r_t^* \Leftrightarrow \sum_{m=1}^{M} p_{mt} x_m^{r_{it}} > \sum_{m=1}^{M} q_{mt} x_m^{r_{it}} \tag{6.28}$$

The interpretation of r_t^* is the risk-attitude parameter that would make a subject exactly indifferent between the two lotteries in task t, and we shall refer to it as the "threshold risk

attitude" associated with task t. An immediate problem is that there is no general closed-form solution for r_t* in (6.28). To progress with estimation of the RP model, we need to restrict attention to special cases.

Consider, for example, the case with only three possible money outcomes, 0, 1, and 2. In this case:

$$\begin{aligned} V_i^{EU}(\mathbf{p_t}; r_i) &= p_{2t} + p_{3t} 2^{r_i} \\ V_i^{EU}(\mathbf{q_t}; r_i) &= q_{2t} + q_{3t} 2^{r_i} \end{aligned} \tag{6.29}$$

If we define $d_{2t} = p_{2t} - q_{2t}$ and $d_{2t} = p_{3t} - q_{3t}$, then the condition for choosing $\mathbf{p_t}$ over $\mathbf{q_t}$ **is:**

$$d_{2t} + d_{3t} 2^{r_{it}} > 0 \tag{6.30}$$

which may be rearranged to give:

$$r_{it} > \frac{\ln\left(-\frac{d_{2t}}{d_{3t}}\right)}{\ln(2)} \equiv r_t^* \tag{6.31}$$

In (6.31), by virtue of the simplicity of the setting, we have a closed form expression for the threshold risk aversion. If the subject's current risk attitude is larger than this threshold, they will choose the riskier lottery $\mathbf{p_t}$. Note that r_t^* is not defined if task *t* is a dominance problem. However, the decision rule for a dominance problem in the RP model is simple: the dominating alternative, $\mathbf{p_t}$ will be chosen with certainty.

For a non-dominance problem, the probability that the riskier lottery is chosen, conditional on the subject's mean risk attitude, m_i, is:

$$\begin{aligned} P(y_{it} = 1 | m_i) &= P(r_{it} > r_t^* | m_i) = P(\ln(r_{it}) > \ln(r_t^*) | m_i) \\ &= P\left(Z > \frac{\ln(r_t^*) - m_i}{\sigma}\right) = \Phi\left(\frac{m_i - \ln(r_t^*)}{\sigma}\right) \end{aligned} \tag{6.32}$$

where $\Phi(.)$ is the standard normal c.d.f. Note that *Z* is being used to represent a standard normal random variable.

If we again define the binary indicator y_{it} =1(-1) if subject i chooses $\mathbf{p_t}$ ($\mathbf{q_t}$), the probabilities of the two choices (conditional on m_i) are then given by:

$$P(y_{it} | m_i) = \Phi\left(y_{it} \times \frac{m_i - \ln(r_t^*)}{\sigma}\right) \tag{6.33}$$

Between-subject heterogeneity is captured by assuming a between-subject distribution for m_i. Because m_i is the mean parameter for a lognormal distribution (6.26), there is no requirement for m_i to be positive, and a natural choice is therefore:

$$m \sim N\left(\mu, \eta^2\right) \tag{6.34}$$

The parameter η^2 represents between-subject variation in risk-attitude. Similarly to the Fechner model in Section 6.8, (6.33) and (6.34) may be combined to constuct a log-likelihood function that can be maximised using MSL.

One further drawback with the RP model is that, as mentioned above, it cannot explain violations of dominance. This is because, if RP is the true model, dominating alternatives will always be chosen. Since many risky choice data sets contain dominance violations, this may appear to be a serious shortcoming of the RP model. However, introducing a tremble parameter, as done in Section 6.5 for the Fechner model, addresses this problem, since dominance violations can be attributed to trembles.

6.10 Non-nested Tests

Having estimated both the Fechner and RP models on the same data set, an obvious question is which model fits the data better. Because the two models are non-nested, we need to consider non-nested tests.

A particular non-nested test that has become popular in these situations is the Vuong Test (Vuong, 1989). In (6.20) we presented the per-subject likelihood contribution for the Fechner model. Let $\hat{f}_i, i = 1, \cdots, n$ be these contributions with parameters replaced by MLEs. Similarly, let $\hat{g}_i, i = 1, \cdots, n$ be the corresponding for the RP model. The Vuong Test test is based on the quantity D, defined by:

$$D = n^{-1/2} \sum_{i=1}^{n} \ln\left(\frac{\hat{f}_i}{\hat{g}_i}\right) \tag{6.35}$$

The quantity D defined in (6.35) is similar to the log-likelihood ratio of the two models, but since the models are non-nested, it can be of either sign. An estimate of the variance of D is given by:

$$\hat{V} = n^{-1} \sum_{i=1}^{n} \left(\left[\ln\left(\frac{\hat{f}_i}{\hat{g}_i}\right) \right]^2 - \left[\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{\hat{f}_i}{\hat{g}_i}\right) \right]^2 \right) \tag{6.36}$$

The Vuong test statistic is given by:

$$Z = \frac{D}{\sqrt{\hat{V}}} \tag{6.37}$$

As proved by Vuong (1989), the test statistic Z in (6.37) has a limiting standard normal distribution under the null hypothesis that the two models are equivalent. A significantly positive value of Z would indicate that the Fechner model is closer to the true data generating process than the RP model. A significantly negative value would indicate the converse.

Loomes et al. (2002) compare Fechner and RP models using Vuong's (1989) non-nested test. They find clear superiority of RP over Fechner, in the context of the choice experiment they work with.

Clarke (2003) considered non-parametric alternatives to the Vuong test. Either the Wilcoxon signed ranks test or the sign test (for both see Section 2.11) can be applied to the paired likelihood contributions. Clarke (2003) argues that there is no reason to expect symmetry to hold in differences of log-likelihoods of competing models, so the sign test is preferred.

7. Discrete Heterogeneity: Finite Mixture Models

Finite Mixture modelling has become very important in Experimetrics, as a way of dealing with discrete heterogeneity, that is, a situation in which the population is made up of a finite set of distinct types. The framework has been applied to a number of areas in Experimental Economics, for example probabilistic updating rules (El-Gamal and Grether, 1995), contributions to public goods (Bardsley and Moffatt, 2007), and Fairness experiments (Cappelen et al., 2007; Conte and Moffatt, 2014).

Another very useful application of mixture modelling is to depth of reasoning models (see e.g. Bosch-Domènech et al., 2010), in which subjects are divided between levels of cognitive ability. This application is used for illustration in this section.

7.1 Finite mixture models

Finite mixture models, or just mixture models, are a class of model that offer a means of separating subjects into different types. A thorough general treatment of this framework is provided by McLachlan and Peel (2000). These models are labelled as "finite" mixture models because a finite number of types is being assumed. An "infinite" mixture model, if such a label were used, would correspond to a random coefficient model, or random effects model, in which it is assumed that there is continuous variable in some parameter indexing behavioural type.

In this section, we will derive the log-likelihood functions for various types of mixture model in a general context. The formulae derived below are the ones that are used in specific contexts in later sections.

There are broadly two approaches to mixture modelling. The first, and the one to which we restrict attention here, amounts to deciding on the number of types, K say, at the outset, and also specifying the "model" applying to each type. The econometric objective is to estimate the parameters of each of the K models, along with the K "mixing proportions" (that is, the proportion of each type in the population). The mixing proportions are denoted as $\pi_1 \ldots \pi_K$. The other approach is to allow the data to determine the number of types, as well as the equations defining them. The resulting models fall under the heading of "latent class models" which have been surveyed by Collins and Lanza (2009). The reason for avoiding these models here is that they tend to identify types for which there is no meaningful label, and hence it can be hard to draw useful conclusions following estimation.

The structure of the log-likelihood function depends on whether the decision variable is being treated as discrete or continuous. Consider first a discrete outcome Y taking a finite number of possible values y^1, y^2 ,…,y^J, each of these values representing one of J possible strategies or actions. Consider a K-type mixture model, with types labelled as T1…TK. Let

$p_k^j \;\; j=1,\cdots J; k=1,\cdots,K$ be the probability (according to the assumed model) that a Lk player chooses action y^j. Then the sample Log-likelihood function, based on an observed sample $y_1,\ldots,y_n$ may be constructed as:

$$LogL = \sum_{i=1}^{n}\sum_{j=1}^{J} I\left(y_i = y^j\right) \ln\left(\sum_{k=1}^{K} \pi_k p_k^j\right) \tag{7.1}$$

where I(.) is the indicator function. If $n^j \;\; j=1,\cdots J$ is the number of subjects choosing value y^j, then (7.1) may be written more simply as:

$$LogL = \sum_{j=1}^{J} n^j \ln\left(\sum_{k=1}^{K} \pi_k p_k^j\right) \tag{7.2}$$

Maximising (7.2) with respect to the K mixing proportions gives maximum likelihood estimates (MLEs) of these parameters.[41] Let the MLEs be $\hat{\pi}_1,\cdots,\hat{\pi}_K$.

Having estimated the model, the posterior type probabilities for a player choosing y^j can be obtained (using Bayes' Rule) as:

$$\hat{P}\left(Tk \mid y^j\right) = \frac{\hat{\pi}_k p_k^j}{\sum_{k'=1}^{K} \hat{\pi}_{k'} p_{k'}^j} \quad k=1,\cdots,K \tag{7.3}$$

Next, consider the continuous case. Again let Y be the decision variable, but now assume it has continuous distribution with density $f(y;\theta)$, where θ is a vector of distributional parameters. We need to define the density functions conditional on each type: $f(y|T1;\theta)$, …, $f(y|TK;\theta)$. Then the sample log-likelihood may be constructed as:

$$LogL = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k f\left(y_i \mid Tk;\theta\right)\right) \tag{7.4}$$

Maximisation of (7.4) gives MLEs of the mixing proportions, and also of the distributional parameters contained in θ.

The posterior type probabilities for the continuous case, for a Player choosing the value y, are obtained, again using Bayes' Rule, as:

$$\hat{P}\left(Tk \mid y\right) = \frac{\hat{\pi}_k f\left(y \mid Lk;\hat{\theta}\right)}{\sum_{k'=1}^{K} \hat{\pi}_{k'} f\left(y \mid Lk';\hat{\theta}\right)} \quad k=1,\cdots,K \tag{7.5}$$

where hats again denote MLEs. The posterior type probabilities defined in (7.3) and (7.5) are useful for categorising players into types. However, note that there is no claim to be able to identify any individual subject as belonging to any particular type with certainty, although, in situations where data is informative, posterior type-probabilities can be very close to one.

[41] Since the mixing proportions sum to 1, (7.2) is maximised with respect to K-1 of them, and the K-th estimate is deduced.

As usual, we also need to consider the case in which there are multiple observations per subject. In the continuous case, let y_{it} be subject i's decision in task t, t=1,…T. Then the Log-likelihood function (7.4) becomes:

$$LogL = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k \left[\prod_{t=1}^{T} f\left(y_{it} \mid Tk; \theta\right) \right] \right) \tag{7.6}$$

Note that there is an assumption implicit in (7.6) that a given subject is of the same type for all tasks. This is another important feature of the mixture modelling framework. If subjects are required to be allowed to switch between types, a different sort of model is required, namely the Markov-switching model (see e.g. Shachat et al., 2015).

7.2 Depth of Reasoning Models

Depth of Reasoning models have become very popular in the analysis of behaviour in interactive games (see Stahl and Wilson, 1995; Nagel, 1995). The key feature of the approach is the assumption that agents divide between different, discrete, levels of reasoning. Each player is assumed to form a belief about the levels of reasoning (and therefore the strategies) of other players, and adopts a strategy which is the best response to these assumed strategies. Of course, in many settings, the majority of players have *incorrect* beliefs about the other players' strategies, and this fact is embraced in the modelling strategy. The approach is most useful for the modelling of initial responses, that is, behaviour observed when a game is played only once, with no opportunity for players to learn what to expect. This is, of course, the situation that applies in many real life applications of game theory such as international conflicts.

Perhaps the most well-known depth of reasoning model is the Level-k model, in which the different levels of reasoning are characterised as follows. A completely non-strategic player will choose actions without regard to the actions of other players. Such a player is said to have zero-order beliefs and may also be called a level-zero player (L0). A slightly more sophisticated (level-one, L1) player believes that all other players will act non-strategically, that is, that all others are L0; his or her action will be the best response consistent with this first-order belief. An even more sophisticated (level-two, L2) player best-responds to the belief that all other players are L1. This pattern continues for higher-level players, but each player has only a finite depth of reasoning, meaning that individual players have a limit to the depth to which they can reason strategically. This is the sense that level-k models lie squarely within in the realms of "bounded rationality" theory. Of course if a player has immaculate powers of reasoning, they may be classified as "level-infinity" (L∞) and their behaviour will correspond to the Nash prediction.

The finite mixture model, explained in Section 7.1, is a natural method for dividing subjects between levels of reasoning. The "types" are simply levels of reasoning. The number of reasoning levels is usually assumed to be quite low, on the basis that we do not expect agents to operate at levels beyond 3 or 4. However, we sometimes include a "level-∞" type whose behaviour corresponds to the Nash equilibrium prediction.

An important issue is what value of K should be chosen. That is, how many types should the mixture contain. This is clearly a model selection issue, since the value of K can be varied and the model that best fits the data can be adopted. When assessing a model's fit to the data, it

is clear that an adjustment needs to be made for the number of parameters, and for this reason Akaike's Information Criterion (AIC) is an appropriate choice of model selection criterion. AIC is given by 2k-2logL where k is the number of parameters in the model. The preferred model is the one with the lowest AIC.

7.3 The 11-20 Money Request Game

The 11-20 Money request game (Arad and Rubinstein, 2012) has two players. Each player requests an amount of money. The amount of money must be a whole number between 11 and 20. Each player receives the amount that he or she requests. However, if one player's request is *exactly one unit less* than that of the other player, the player requesting the smaller amount receives a bonus of 20. The game has a mixed strategy Nash equilibrium covering requests between 16 and 20.[42]

This game is particularly well-suited to studying level-k reasoning, for the following reasons. First, the maximal request of 20, being the instinctive choice, provides a natural anchor for the iterative reasoning process. It is accordingly assumed that L0 players choose 20. Second, best-responding is straighforward and likely to be error-free: for a L1 player, it is obvious that the best response to L0 play is 19; for a L2 player, it is obvious that the best response to L1 play is 18; and so on. Third, level-k reasoning is unlikely to be distorted by social preferences. This is because, if a player ever receives the bonus of 20, it is not at the expense of the other player.

Arad and Rubinstein (2012) played the version of the game as described above on 108 subjects. The distribution of requests is shown in Table 1. As expected, the majority of requests are towards the top of the 11-20 range. A small number of requests appear in the lower part of the range.

request	11	12	13	14	15	16	17	18	19	20
Frequency (%)	4	0	3	6	1	6	32	30	12	6

Table 7.1: Distribution of requests in the 11-20 Game (n=108). Extracted from Arad and Rubinstein (2012) Table 1.

Apart from asking players for their requests, players were also asked to provide a written explanation for their decision. These explanations were very useful in verifying the use of level-k reasoning.[43] An important point stressed by Arad and Rubinstein (2012) is that no players who requested less than 16 provided an explanation that suggested a level-k reasoning process. Instead, these (14) players making low requests tended to provide explanations suggesting that their choice was essentially random. In view of this, it was considered sensible to add a "random" type to the level-k model. The definition of a "random" type is a player who chooses between the 10 possible requests, 11-20, with equal probability 1/10.

Let us consider estimation of level-k models using the data presented in Table 1, following the approach of Moffatt (2020). We will start by assuming K=4. That is, we assume there are six

[42] According to Arad and Rubinstein (p.3545, top row of Table 1), the mixed strategy Nash equilibrium consists of: 15 with prob 0.25; 16 with prob 0.25; 17 with prob 0.20; 18 with prob 0.15; 19 with prob 0.10; 20 with prob 0.05.

[43] For example, it appears that one player identified themselves clearly as a L2 reasoner by explaining that "I request 18 shekels since many people believe that the majority will request 20 and thus they will request 19" (Arad and Rubinstein, 2012, p.3566).

types in total: L0, L1, L2, L3, L4, and "random". The mixing proportions will be π_0, π_1, π_2, π_3, π_4, and π_r respectively.

To construct the log-likelihood function, we need to consider the probability of each request. The request of 20 occurs if the player is L0, but it also occurs with probability 1/10 if the player is "random". Hence the probability of a request of 20 is:

$$P(request = 20) = \pi_0 + \frac{\pi_r}{10} \tag{7.7}$$

The probabilities of requests of 19, 18, 17 and 16 are obtained similarly to (7.7), since these requests correspond to L1, L2, L3 and L4 respectively. Since we are assuming no types above L4, any request below 16 can only be explained using the random type. Hence each request below 16 has probability $\pi_r/10$.

Using the probabilities just derived, together with the frequencies shown in Table 7.1, we obtain the following sample log-likelihood function:

$$\begin{aligned} LogL(\pi_1,\pi_2,\pi_3,\pi_4,\pi_r) = 6\ln\left((1-\pi_1-\pi_2-\pi_3-\pi_4-\pi_r)+\frac{\pi_r}{10}\right)+12\ln\left(\pi_1+\frac{\pi_r}{10}\right) \\ +30\ln\left(\pi_2+\frac{\pi_r}{10}\right)+32\ln\left(\pi_3+\frac{\pi_r}{10}\right)+6\ln\left(\pi_4+\frac{\pi_r}{10}\right)+14\ln\left(\frac{\pi_r}{10}\right) \end{aligned} \tag{7.8}$$

The Log-likelihood function (7.8) is maximised with respect to the five mixing proportions $\pi_1,\pi_2,\pi_3,\pi_4,\pi_r$. The results are shown in the "K=4" column of Table 7.2.

	K=4	K=3	K=3 (no L0)
π_0	0.037(0.025)	0.031(0.025)	
π_1	0.093**(0.033)	0.086**(0.033)	0.082*(0.033)
π_2	0.269**(0.046)	0.262**(0.046)	0.258**(0.047)
π_3	0.287**(0.050)	0.281**(0.047)	0.276**(0.048)
π_4	0.037(0.025)		
π_r	0.278**(0.067)	0.340**(0.065)	0.384**(0.061)
n	108	108	108
LogL	-197.80	-199.32	-200.30
AIC(=2k-2logL)	407.60	408.64	408.60

Table 7.2: MLEs of parameters of three Level-k models applied to data from the 11-20 Money Request Game. Asymptotic standard errors in parentheses. The preferred model is the one with the lowest AIC.

We see from the "K=4" column of Table 2 that the majority of subjects divide roughly equally between L2, L3 and "random". In fact, the mixing proportions π_0 and π_4 are insignificantly different from zero. This suggests estimating models without these types. The column "K=3" shows results of the model with L4 excluded. This model has a noticeably higher estimate of π_r, indicating that players previously recognized as L4 are now recognized as random. Akaike's Information Criterion (AIC) indicates that model "K=3" is inferior to model "K=4". Removing L0 as well (final column) brings about a further rise in π_r, but this model is also found to be inferior to model "K=4" according to AIC.

We are led to conclude that the best model for explaining this data set is a level-k model including all levels from 0 to 4, and also a random type. We are also led to conclude that the majority of players divide between L2, L3 and "random".

7.4 Guessing games

Nagel's (1995) guessing game, sometimes called the "ω-Beauty Contest game"[44], takes the following form. Each player chooses a whole number between 0 and 100. The winner (with the usual choice of ω=2/3) is the player whose number is closest to 2/3 of the average for the entire group.

This is another setting in which the level-k model is a highly appropriate framework for describing behaviour. Level-0 reasoners are assumed to select a number completely at random, from a (discrete) uniform distribution on [0,100]. Level-1 reasoners are assumed to believe that all other players are level-0 reasoners, inferring that the mean guess will be around 50, and therefore that the best guess is 33 (being the closest integer to 2/3 of 50). Level-2 reasoners are assumed to believe that all others are Level-1, with a mean guess of 33, so that the best guess for this type of individual is 22. This sequence continues. Level-3 reasoners will guess 15. Level-4 reasoners will guess 10, and so on.

Note that if every player had immaculate powers of reasoning (i.e. L∞), and believed that all other players had the same powers, they would all supply a guess of 0, and they would all be correct and share the prize. This is the Nash Equlilibrium prediction for the guessing game. However, needless to say, this is not what happens when the game is played with real subjects.

Bosch-Domènech et al. (2010) collected a large amount of data on guesses in beauty contest games played by newspapers with their readers as contestants. We will work with these data in this section. The distribution of guesses is shown in Figure 7.1. We see that the distribution is multi-modal, with one clear mode at around 33 and another around 22. Note that there is another mode close to zero. These modes will be explained using the level-k model.

[44] Although the Beauty Contest Game was first introduced to the Experimental Economics literature by Nagel (1995), it appears to have been invented in a popular computing magazine in the late 1970's. See Nagel et al. (2017) for an inspiring account of the history of the game. The name "Beauty Contest Game" derives from a paragraph on p. 156 of Keynes' (1936), *General Theory of Employment, Interest and Money*, in which the problem of choosing stocks for a portolio is likened to "those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole". In essence, the competition described here by Keynes is a guessing game with ω=1.

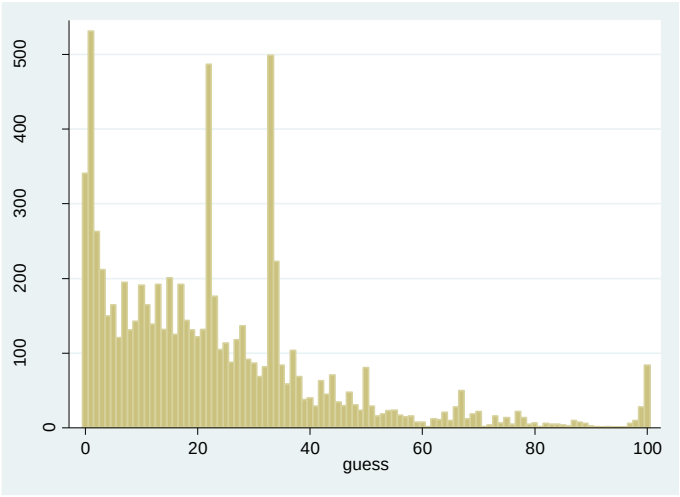

Figure 7.1: The distribution of guesses from the newspaper guessing games; collected by Bosch-Domènech et al. (2010). Sample size: 7,892.

The estimation problem is to use the data shown in the histogram of Figure 1 to estimate the proportion of the population who are at each level of reasoning. We will once again assume that there are a finite number K+1 of types, with the maximum level of reasoning K.

Apart from Level-0 reasoners, who are assumed to choose from a (discrete) uniform distribution, we assume that each player's choice is the best guess for a player of their reasoning level, plus a random normally distributed error with mean zero and standard deviation σ. That is, we assume that if y_k^* is the best guess for Lk, then the actual guess y will be determined by (following Bosch-Domènech et al., 2010; Runco, 2013; Moffatt, 2020):

$$(y \mid Lk) = y_k^* + \varepsilon \quad \varepsilon \sim N\left(0, \sigma_k^2\right) \quad k = 1, \cdots, K \tag{7.9}$$

In (7.9) it is assumed that observed guesses are random departures (due to computational error) from the best guess, with σ_k^2, $k = 1, \cdots, K$ representing the extent of computational errors made by a player of Lk. If we set K=4, the "best guesses" are y_1^*=33.3, y_2^*=22.2, y_3^*=14.8, y_4^*=9.9. For a Nash (i.e. L∞) type, we need to allow for the fact that the Nash prediction is at the limit of the range of possible guesses. Hence we assume a censored normal distribution for the Nash type, as follows:

$$(y \mid L_{Nash}) = \varepsilon \quad \varepsilon \sim CN\left(0, \sigma_{Nash}^2\right) \tag{7.10}$$

where $CN\left(0, \sigma_{Nash}^2\right)$ represents a distribution consisting of a probability mass of one half at zero, and a half normal density above zero. Strictly speaking, we should also allow for censoring at zero (and perhaps 100 also) for other reasoning levels. However, the best guesses for most reasoning levels are typically sufficiently far from the limits for censoring to make any difference in estimation.

The distributional assumptions (7.9) and (7.10) give us the conditional density/probability functions for each type:

$$
\begin{aligned}
f(y \mid L_0) &= 1/100 \qquad 0 \le y \le 100 \\
f(y \mid L_k) &= \frac{1}{\sigma_k}\phi\left(\frac{y - y_k^*}{\sigma_k}\right) \quad 0 \le y \le 100 \;\; k = 1, \cdots, K \\
f(y \mid L_{Nash}) &= \frac{1}{2} I(y = 0) + \frac{1}{\sigma_{Nash}}\phi\left(\frac{y}{\sigma_{Nash}}\right) I(y > 0)
\end{aligned}
\tag{7.11}
$$

where $I(.)$ is the indicator function. Combining the the conditional densities (7.11) with the mixing proportions $\pi_0, \pi_1, \cdots, \pi_K, \pi_{Nash}$, gives the sample log-likelihood (for a sample of guesses $y_i, i = 1, \cdots, n$):

$$
LogL = \sum_{i=1}^{n} \ln \left[\begin{array}{l} \frac{\pi_0}{100} + \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}\phi\left(\frac{y_i - y_k^*}{\sigma_k}\right) \\ + \pi_{Nash}\left(\frac{1}{2} I(y_i = 0) + \frac{1}{\sigma_{Nash}}\phi\left(\frac{y_i}{\sigma_{Nash}}\right) I(y_i > 0)\right) \end{array} \right] \tag{7.12}
$$

Maximisation of (16) gives MLEs of the mixing proportions and also the variance parameters.

Estimation has been performed on the data set of Bosch-Domènech et al. (2010), for various values of K, by Moffatt (2020) and the results are reproduced in Table 7.3.

	K=1	K=2	K=3	K=4
π_0	0.138(0.008)	0.169(0.007)	0.172(0.007)	0.173(0.007)
π_1	0.586(0.013)	0.062(0.003)	0.073(0.003)	0.070(0.003)
π_2		0.634(0.010)	0.355(0.018)	0.441(0.021)
π_3			0.291(0.018)	0.061(0.039)
π_4				0.151(0.026)
π_{Nash}	0.277(0.009)	0.136(0.006)	0.109(0.006)	0.103(0.006)
σ_1	14.676(0.279)	0.459(0.018)	0.482(0.020)	0.475(0.020)
σ_2		12.457(0.199)	13.978(0.352)	12.536(0.367)
σ_3			8.375(0.267)	7.994(0.610)
σ_4				6.299(0.322)
σ_{Nash}	7.276(0.337)	2.452(0.132)	1.986(0.102)	1.861(0.096)
n	7,892	7,892	7,892	7,892
LogL	-33,859.76	-32,424.59	-32,242.84	-32,217.31
AIC (=2k-2logL)	67,727.52	64,859.18	64,497.68	64,448.62

Table 7.3: MLEs of parameters of various Level-k models applied to Bosch-Domènech et al.'s (2010) Guessing Game data. Results previously reported by Moffatt (2020). The preferred model is the one with the lowest AIC. KS is the Kolmogorov-Smirnov test, and the numbers in this row are KS test statistics.

We again use AIC to choose between models. Here it seems that the model fit improves every time a new level of reasoning is added. We will therefore interpret results from the model with the highest number of levels, K=4. We see that the modal level appears to be L2, with 44% of players estimated to be of this type. We also see that around 17% of players are L0,

while around 10% are "Nash". The dispersion parameter appears to be very low for L1, very high for L2, and then to decrease when level rises beyond this.

Post-estimation, we may generate posterior type probabilities using Bayes' rule (7.5). These are plotted against the subject's guess in Figure 7.2. The dashed curve represents the posterior probability of Level-0. Note that this is close to 1 for any subject whose guess is greater than about 60. The other posterior probabilities peak in different positions, as expected. The curve peaking at 33 is the level-1 posterior probability; the one peaking at a slightly lower value but with much wider dispersion is that for level-2; the one peaking at around 10 is for level-3; the one at 5 is for level 4. The curve peaking at zero (or one) is for the "Nash" type. The position of this last curve indicates that subjects whose guess is zero *or a very small positive number* may be categorised as "Nash".

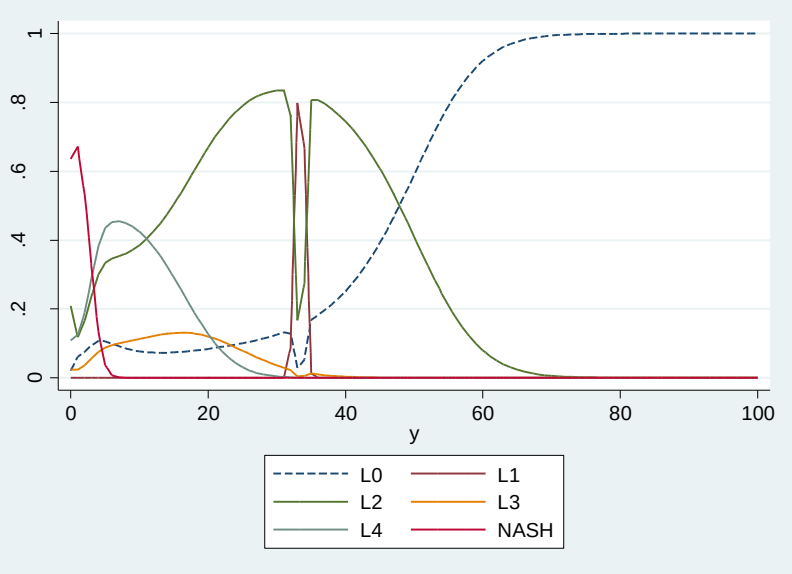

Figure 7.2: Posterior type probabilities in the level-k model for the Guessing Game.

7.5 Other Depth of Reasoning Models

A shortcoming of the version of the level-k model constucted and estimated in the last section is that it assumes that all players believe that all other players are exactly one level below themselves. An obvious reason to doubt this assumption is that if a player possesses the cognitive ability associated with level 2 or above, they are surely not naïve enough to believe that all other subjects are only one level below.

The Cognitive Hierarchy (CH) model, developed by Camerer at al. (2004), addresses this issue by assuming that the distribution of the population over reasoning levels is Poisson (τ), and that a player of type k believes other members of the population to be distributed between types 0…(k-1) according to an upper truncated Poisson (τ) distribution. That is, the (true) distribution of types over the population is given by:

$$p(k) = \left(\frac{e^{-\tau}\tau^{k}}{k!} \right) \quad k = 0,1,2,\cdots \tag{7.13}$$

and a player of type k believes that the probability distribution of other subjects between types is:

$$p_k(j)=\left(\frac{e^{-\tau}\tau^j}{j!}\right)\Bigg/\left(\sum_{m=0}^{k-1}\frac{e^{-\tau}\tau^m}{m!}\right)\quad j=0,\cdots,k-1 \tag{7.14}$$

On the basis of these probabilities, the "best guesses" for each type may be computed recursively as:

$$\begin{aligned}
&b_1=0.67\times 0.5=0.33\\
&b_2=0.67\left[p_2(1)\times b_1+p_2(0)\times 0.5\right]\\
&b_3=0.67\left[p_3(2)\times b_2+p_3(1)\times b_1+p_3(0)\times 0.5\right]\\
&b_4=0.67\left[p_4(3)\times b_3+p_4(2)\times b_2+p_4(1)\times b_1+p_4(0)\times 0.5\right]\\
&\vdots
\end{aligned} \tag{7.15}$$

As in the level-k model, we also assume the existence of a Nash type whose best guess is zero. This is achieved by setting a value of K and assuming that all players with level greater than K are Nash types. The proportion of players who are Nash types is then $1-\sum_{k=0}^{K}p(k)$.

The Log-likelihood function is then constructed by combining the observed guesses (*y*) with the type probabilities (*p(k)* given by (7.5) above) and the best guesses, b_k:

$$LogL=\sum_{i=1}^{n}\ln\left[\begin{array}{l}\dfrac{p(0)}{101}+\sum_{k=1}^{K}p(k)\dfrac{1}{\sigma}\phi\left(\dfrac{y_i-b_k}{\sigma}\right)\\ +\left(1-\sum_{k=0}^{K}p(k)\right)\left(\dfrac{1}{2}I(y_i=0)+\dfrac{1}{\sigma_{Nash}}\phi\left(\dfrac{y_i}{\sigma_{Nash}}\right)I(y_i>0)\right)\end{array}\right] \tag{7.16}$$

Note that there are only two parameters to estimate in this model: the computational error parameter (σ) and the Poisson mean (τ).

Moffatt (2020) has applied the cognitive hierarchy model (7.16) to the data of Bosch-Domènech et al. (2010). The Poisson mean (τ) is estimated as 2.75 with a small standard error. The interpretation is that the mean reasoning level over the population of players is 2.75. This seems rather high in the light of Camerer et al.'s (2003) assertion that this parameter is usually around 1.5. Moffatt (2020) finds that the model fit of the CH model, estimated as in (7.16) is inferior to that of the level-k models estimated in the last section. However, if the dispersion parameter (σ) were allowed to vary between levels, as done in the level-k model, this would be likely to improve the fit of the CH model.

Other ways of improving the fit of the CH model have been explored. Chong et al. (2016) consider a generalised cognitive hierarchy model that nests CH and level-k models. Stahl and Wilson (1995) extend the CH model to allow for a rational expectations (RE) type. An RE-player is one who realises that the population is divided between lower level types and Nash types, but also realises that there exist other RE types. In the context they consider, they do not find evidence of the existence of RE.

Moffatt et al. (2020), introduce a "sophisticated" type to the level-k model. A "sophisticated" player is similar to Stahl and Wilson's (1995) RE type, but, although a sophisticated player recognises the presence of other sophisticated players in the population, they do not necessarily hold correct beliefs about the proportion of such players. In fact, this turns out to be important. Moffatt et al. (2020) find that the proportion of sophisticated agents in the population is quite small (around 10%), but these sophisticated players tend to believe that this proportion is much higher than this, and this incorrect belief leads to guesses that are too low to win the prize. The discrepancy of beliefs is interpreted as a manifestation of the Dunning-Kruger effect,[45] a behavioural anomaly that predicts that experts overestimate the ability of those around them.

7.6 Other Applications of the Finite Mixture Model

Depth of reasoning models were relatively simple to estimate because the types were clearly identified by the best responses which come directly from the theoretical model.

Some other types of finite mixture model are more complicated, in the sense that types are defined by regression equations. For example Bardsley and Moffatt (2007) estimate a 4-type mixture model with data from a public goods experiment. Type 1 is the free-rider (mixing proportion: 25%) who tends to contribute zero in every round regardless of others' contributions or their own position in the sequence. Type 2 is the reciprocator (30%) whose contribution depends positively on the contributions of others placed earlier in the sequence. Type 3 is the selfish contributor (39%) who contributes in anticipation of reciprocity by others, and whose contribution therefore depends negatively on position in the sequence. Type 4 is the altruist (6%) whose contribution is positive and does not depend on others' contributions or their own position in the sequence.

Conte and Moffatt (2014) consider data from a fairness experment, in which pairs of players engage in an effort task to earn a pie, and then decide how the pie should be divided between them. Although the results are sensitive to the choice of econometric specification, a robust finding is that the population divides roughly equally between those who consider the source of the endowment relevant to the decision of the final allocation, and those who do not.

7.7 Machine Learning Models

Very recently, machine Learning Models have gained popularity as a means of dividing subjects into types.

Fallucchi et al. (2020) reports on the analysis of two previously published contest experiments, one with fixed matching, the other with random matching. The novel feature of the analysis is the use of the classifier-Lasso (C-Lasso) method (Su et al., 2016) extended to the Tobit model. We will summarise this estimation framework here.

First consider the following version of the Tobit model:

$$y_{it} = \max\left(0, \mu_i + \mathbf{x}_{it}'\boldsymbol{\beta}_{\mathbf{i}} + \varepsilon_{it}\right) \quad \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right) \; i = 1, \cdots, N \; t = 1, \cdots, T \tag{7.17}$$

[45] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

In (7.17) y_{it} is the effort of subject i in task t, and $\mathbf{x_{it}}$ is a vector of explanatory variables. Note that all subjects have a subject-specific intercept (μ_i) and a subject-specific slope vector ($\boldsymbol{\beta}_i$).

The purpose of the C-Lasso method is to shrink the set of N subject-specific coefficient vectors $\boldsymbol{\beta}_i$ to a smaller set of K<N group-specific coefficients $\boldsymbol{\beta}_k$:

$$\begin{aligned} y_{it} &= \max\left(0, \mu_i + \mathbf{x_{it}}'\boldsymbol{\beta}_k + \varepsilon_{it}\right) \quad \boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k'} \Leftrightarrow k \neq k' \\ \varepsilon_{it} &\sim N\left(0, \sigma_\varepsilon^2\right) \quad k = 1, \cdots, K \quad i = 1, \cdots, N \quad t = 1, \cdots, T \end{aligned} \tag{7.18}$$

This is achieved by minimising the penalised nonlinear likelihood (PNL) function:

$$\min_{\mu_i, \beta_i, \beta_k, \sigma_\varepsilon} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Psi\left(y_{it}, x_{it}; \boldsymbol{\beta}_i, \sigma_\varepsilon\right) + \frac{\lambda}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_k \right\| \tag{7.19}$$

The first term in (7.19) represents the log-likelihood function for the Tobit model, and the second term is a penalisation term. The parameter λ acts as a shrinkage parameter on the $\boldsymbol{\beta}_i$ coefficients. If λ=0, the model represents maximal heterogeneity, while a large value of λ leads to perfect homogeneity. Hence the minimisation (7.19) is accompanied by an optimisation over λ. The choice of λ addresses the overfitting/underfitting tradeoff.

Fallucchi et al. (2020) consider an experiment with fixed matching and another with random matching. For each, they obtain a coefficient vector for each subject, and use the C-Lasso method just outlined to shrink the set of vectors to just three, corresponding to "Reciprocators" (58%), "Gamesmen" (9%), and "Others" (32%). Their most significant finding is that the estimated proportion of reciprocators is much lower under random matching than under fixed matching.

A similar approach is followed by Engel (2020), who adopts the Classification and Regression Tree (CART) framwork to address the problem of patterned heterogeneity in experimental data. The method consists of two steps: estimation of a linear regression separately by subject; then assignment of subjects to types using the coefficients in these regressions. The optimal partition of subjects into types is found using machine learning techniques. Engel (2020) claims that the method overcomes problems associated with ml estimation of finite mixture models: non-convergence in likelihood maximisation; the need to deal with dependence; and the need to fix the number of types before estimation.

Fallucchi et al. (2019) use hierarchical clustering analysis to identify types in public goods voluntary contribution games. They construct a typology of behaviour based on a similarity measure. They identify four types with distinct stereotypical behaviours, which together account for about 90% of participants. Compared to the previous approaches, their method produces a classifcation in which different types are more clearly distinguished in terms of strategic behaviour and the resulting economic implications.

Penczynski (2019) uses a combination of Natural Language Processing and Machine Learning to infer levels of reasoning in games. The data consist of text from transcripts of communications between team members. He finds that Machine Learning models of classification closely match out-of-sample the human classification that was previously the norm. This is promising because it opens the way for the reliable analysis of large-scale textual data sets for similar purposes.

8. Other models for behaviour in games

In Section 7 we were mainly interested in the analysis of data from one-shot games. Here we progress to consider methods for modelling data from repeated games. Hence we are shifting the focus from the "short run" to the "long run", and we are mainly interested in the extent to which long-run behaviour matches theoretical predictions.

We start this section by considering simple non-parametric procedures for testing adherence to the mixed strategy equilibrium in repeated games, and for testing randomness in players' choices. Then we progress to parametric models. The focal point of this section is the Quantal Response Equilibrium (QRE) model, which is an econometric model built on the assumption that agents depart from the Nash prediction in a random manner. The basic version of the model contains only one parameter, μ, representing the amount of "noise" in subjects' decisions. The estimation problem is non-standard because, while for standard models, choice probabilities can be computed in closed form for given parameter values, for this problem computation of the choice probabilities for any given value of μ requires the use of non-linear optimisation routine. Hence this non-linear optimisation routine is being called within the function evaluation program.

8.1 Modelling choices in repeated games

We shall use the "pursue-evade" game as an example, with data from Rosenthal et al. (2003). Subjects are arranged in pairs. In each pair, one player takes the role of "Pursuer", the other of "Evader". The game is then played many times. In each repetition of the game, each player has two choices, Left and Right. If they choose differently, the Evader has successfully evaded, and no money is transferred. If they choose the same, the Pursuer has "found" the Evader, and an amount of money is transferred from the Evader to the Pursuer. The amount transferred depends on where the choices coincide: if they both choose Left, one unit is transferred; if they both choose Right, two units are transferred.

The pay-off matrix for the pursue-evade game as just described is as follows:

Pursuer	Evader		
		Left	Right
	Left	1,-1	0,0
	Right	0,0	2,-2

There is a mixed-strategy Nash equilibrium which is solved as follows.[46] The key is to find a pair of strategies that make both players *indifferent* between the two choices. Suppose Pursuer plays L with probability p_{PL} and R with probability $1-p_{PL}$, while Evader plays L with probability p_{EL} and R with probability $1-p_{EL}$. Then, for Pursuer, the expected values from playing L and R are (in self-explanatory notation):

$$\begin{aligned} EV_P(L) &= p_{EL}\times 1+(1-p_{EL})\times 0 = p_{EL} \\ EV_P(R) &= p_{EL}\times 0+(1-p_{EL})\times 2 = 2(1-p_{EL}) \end{aligned} \tag{8.1}$$

Hence Pursuer is indifferent between the two choices if:

[46] The method for solving for a mixed-strategy Nash equilibrium is explained well in Appendix A1.1 of Camerer (2003).

$$p_{EL} = 2\left(1 - p_{EL}\right) \Rightarrow p_{EL} = \frac{2}{3} \tag{8.2}$$

For Evader, the expected values from playing L and R are:

$$\begin{aligned} EV_E\left(L\right) &= p_{PL} \times \left(-1\right) + \left(1 - p_{PL}\right) \times 0 = -p_{PL} \\ EV_E\left(R\right) &= p_{PL} \times 0 + \left(1 - p_{PL}\right) \times \left(-2\right) = -2\left(1 - p_{PL}\right) \end{aligned} \tag{8.3}$$

Hence Evader is indifferent if:

$$-p_{PL} = -2\left(1 - p_{PL}\right) \Rightarrow p_{PL} = \frac{2}{3} \tag{8.4}$$

The mixed-strategy Nash equilibrium is therefore characterised by the pair of best responses:

$$\left[\left(\tfrac{2}{3}L, \tfrac{1}{3}R\right), \left(\tfrac{2}{3}L, \tfrac{1}{3}R\right)\right] \tag{8.5}$$

Therefore, in the pursue-evade game described above, we might expect both players to randomise between L and R in such a way that in the long-run, two-thirds of their decisions are L.

8.2 Non-parametric tests on repeated game data

The data used for demonstration purposes is from Rosenthal et al. (2003). In their experiment, the pursue-evade game described exactly as above is played by the 14 pairs, 21-34. Each pair plays the game 100 times. Table 8.1 presents summary statistics for the choices made by the players in these pairs. Analysis of this data set in the ways described in this section has already been reported by Moffatt (2015).

Perhaps the most obvious thing to check is how closely each player conforms to the mixed-strategy Nash equilibrium derived above. We do this by simply observing how close their proportion of "L" choices (shown in columns 2 and 3 of Table 8.1) is to the prediction of 0.67. We see that some of the evaders are remarkably close to the Nash prediction of 0.67. We also see that pursuers are typically below the Nash prediction of 0.67 (i.e. they do not choose L frequently enough).

To test formally for a subject's adherence to the Nash prediction, the null hypothesis is that the proportion of the subject's "left" choices equals 0.6667, and the appropriate test is the binomial test. The p-values for these tests are shown in columns 4 and 5 of Table 8.1, with stars indicating significant departures from Nash. On the evidence of this test, we see that roughly half of all players (whether Pursuer or Evader) have choices that are consistent with the Nash prediction.

A second theoretical prediction that can be tested is randomness in the player's sequence of choices. Clearly, any pattern in the sequence of choices could potentially be exploited by the opponent, so the optimal strategy is to produce as random a sequence as possible.

Consider the choices of the players in pair 21 of Rosenthal et al. (2003) (1 indicates "L"):

Pursuer in pair 21:

1001000100010000100100101010101100010100101100100101011010111000010100100010110001001100111011110010

Evader in pair 21:

1110100011011111111101111111011110111111011111101111110111111101111111011111110111111101111111111111111101111101111111

Pair	Pur_L	Eva_L	Binomial(Pur)	Binomial(Eva)	Runs(Pur)	Runs(Eva)
21	.43	.84	0.000**	0.000**	0.01*	0.67
22	.62	.59	0.340	0.112	0.05	0.74
23	.55	.59	0.015*	0.112	0.36	0.02*
24	.76	.78	0.056	0.019*	0.00**	0.09
25	.59	.86	0.112	0.000**	0.13	0.97
26	.66	.82	0.916	0.001**	0.00**	0.61
27	.53	.67	0.006**	1.000	0.01*	0.69
28	.62	.70	0.340	0.525	0.54	0.63
29	.67	.78	1.000	0.019*	0.10	0.62
30	.53	.59	0.006**	0.112	0.01*	0.90
31	.55	.69	0.015*	0.672	0.76	0.96
32	.56	.69	0.026*	0.672	0.09	0.01*
33	.42	.55	0.000**	0.015*	0.79	0.03*
34	.46	.66	0.000**	0.916	0.34	0.05

Table 8.1: proportions of L-choices in 100 plays of the Pursue-Evade game of Rosenthal et al. (2003); p-values from binomial tests; p-values from runs tests.

In order to test whether these two sequences of 100 numbers are random, we consider the number of "runs", that is, the number of groups of consecutive digits that are the same. The number of runs in the two sequences shown above are 62 and 29 respectively. A natural test of randomness would be one that asks whether the number of runs is approximately what is expected if the sequence is indeed random. If the number of runs is too high, this would imply negative serial correlation (switching too often), while too few runs would imply positive serial correlation (not switching often enough).

The test which does this is the runs test (see Siegel and Castellan, 1988). The null hypothesis implicit in this test is that the sequence of choices is random, and also that the underlying mixed strategy (i.e. the probability of L) is fixed over time. Let m be the number of L-choices and n be the number of R-choices, in a sequence of N=m+n decisions. Let r be the number of runs in the sequence. Provided both m and n are larger than about 20, a good approximation to the sampling distribution of r, under the null hypothesis of randomness of the sequence, is:

$$r \sim N\left[\left(\frac{2mn}{N}+1\right), \frac{2mn(2mn-N)}{N^2(N-1)}\right] \tag{8.6}$$

It follows that the null hypothesis of randomness may be tested using the statistic:

$$z = \frac{r - \left(\frac{2mn}{N} + 1 \right)}{\sqrt{\frac{2mn(2mn - N)}{N^2(N-1)}}} \tag{8.7}$$

Values of z obtained using (8.7) are approximately standard normal under the null hypothesis.

Applying (8.7) to the sequences shown above for pursuer and evader in pair 21, we obtain 2.46 and 0.42 respectively. These results provide evidence that pursuer is switching too often while evader is switching about the right number of times to appear random. Results of the runs test for all players are shown in the final two columns of Table 8.1. While some players appear to violate randomness, the majority appear to produce apparently random sequences.

It has been stressed that the runs test is a test of not only randomness but also of the constancy over time of the mixed strategy. It is conceivable that a random sequence of choices appears non-random if the underlying mixed strategy is changing over time. Shachat et al. (2012) and Ansari et al. (2012) estimate hidden Markov models in which players are assumed to switch between pure and mixed strategies. Their estimation results lead to the conclusions that there are significant amounts of both pure and mixed strategy play, and that there are low transition probabilities between mixed and pure strategies.

8.3 Quantal Response Equilibrium (QRE): Theory

The Quantal Response Equilibrium (QRE) model was developed by McKelvey and Palfrey (1995). It is based on the idea that best responses are not played with certainty. Each player is assumed to calculate the expected value of each of her available actions given her (correct, in equilibrium) belief about her opponent's choice probabilities, and she attempts to respond optimally, but makes random errors in the process. Thus QRE replaces the perfectly rational expectations equilibrium embodied in Nash equilibrium with an imperfect, or "noisy", rational expectations equilibrium. The principle of an equilibrium is maintained by assuming that players estimate expected payoffs in an unbiased way.

The QRE model follows in the tradition of early models of individual choice behavior (Luce, 1959). The key parameter is the "noise parameter", μ. $\mu=0$ indicates no error, and $\mu=\infty$ indicates "all error". Let us return to the pursue-evade game considered in Section 8.1. To derive the QRE for this game, we first need to apply a stochastic term to each of the expected values:

$$\begin{aligned} EV_P^*(L) &= p_{EL} + \varepsilon_{PL} \\ EV_P^*(R) &= 2(1 - p_{EL}) + \varepsilon_{PR} \end{aligned} \tag{8.8}$$

$$\begin{aligned} EV_E^*(L) &= -p_{PL} + \varepsilon_{EL} \\ EV_E^*(R) &= -2(1 - p_{PL}) + \varepsilon_{ER} \end{aligned} \tag{8.9}$$

For convenience, we assume that each of the stochastic terms are independently distributed with the type-I extreme value distribution with scale parameter μ, whose cdf is given by:

$$F(\varepsilon;\mu)=\exp\left(-\exp\left(-\frac{\varepsilon}{\mu}\right)\right) \tag{8.10}$$

With this distributional assumption, together with the assumption that each player selects the alternative with the higher (stochastic) expected value, it is possible to derive the following choice probabilities for the two players (see Maddala, 1983, for further details of this derivation):

$$\begin{aligned} p_{PL} &= \frac{\exp\left(\frac{p_{EL}}{\mu}\right)}{\exp\left(\frac{p_{EL}}{\mu}\right)+\exp\left(\frac{2(1-p_{EL})}{\mu}\right)} \\ p_{EL} &= \frac{\exp\left(-\frac{p_{PL}}{\mu}\right)}{\exp\left(-\frac{p_{PL}}{\mu}\right)+\exp\left(-\frac{2(1-p_{PL})}{\mu}\right)} \end{aligned} \tag{8.11}$$

From these probability formulae, the role of the parameter μ, as the extent of "noise" in decision-making, is clear: when μ is close to zero, the choice probabilities are close to those dictated by the mixed strategy Nash equilibrium; when μ is large, the probabilities become 0.5, meaning that players divide their choices equally between L and R.

8.4 Computing the probabilities in the QRE model

To compute the probabilities of each player's choices, it is required to solve the two equations (8.11) simultaneously for the two unknowns p_{PL} and p_{EL}, for any given value of μ. Since the two equations are clearly non-linear in p_{PL} and p_{EL}, this is a numerical problem. An approach suggested by Moffatt (2015) is to find the values of p_{PL} and p_{EL} that simultaneously minimize the two quantities:

$$\left(p_{PL}-\frac{\exp\left(\frac{p_{EL}}{\mu}\right)}{\exp\left(\frac{p_{EL}}{\mu}\right)+\exp\left(\frac{2(1-p_{EL})}{\mu}\right)}\right)^2 \text{ and } \left(p_{EL}-\frac{\exp\left(-\frac{p_{PL}}{\mu}\right)}{\exp\left(-\frac{p_{PL}}{\mu}\right)+\exp\left(-\frac{2(1-p_{PL})}{\mu}\right)}\right)^2$$

Both of these quantities have a minimum of zero, and the unique pair values of p_{PL} and p_{EL} that make them both equal to zero will be the required values of p_{PL} and p_{EL}.

Using a suitable optimising algorithm,[47] these probabilities can be computed for a range of values of μ. The resulting probabilities are plotted against μ in Figure 8.1. As expected both probabilities are 0.67 when μ is close to zero. This corresponds to the mixed-strategy Nash equilibrium with zero noise. As the noise-level increases both probabilities gradually approach 0.5, as expected.

[47] The optimiser used by Moffatt (2015) is the "Optimize" command in Mata.

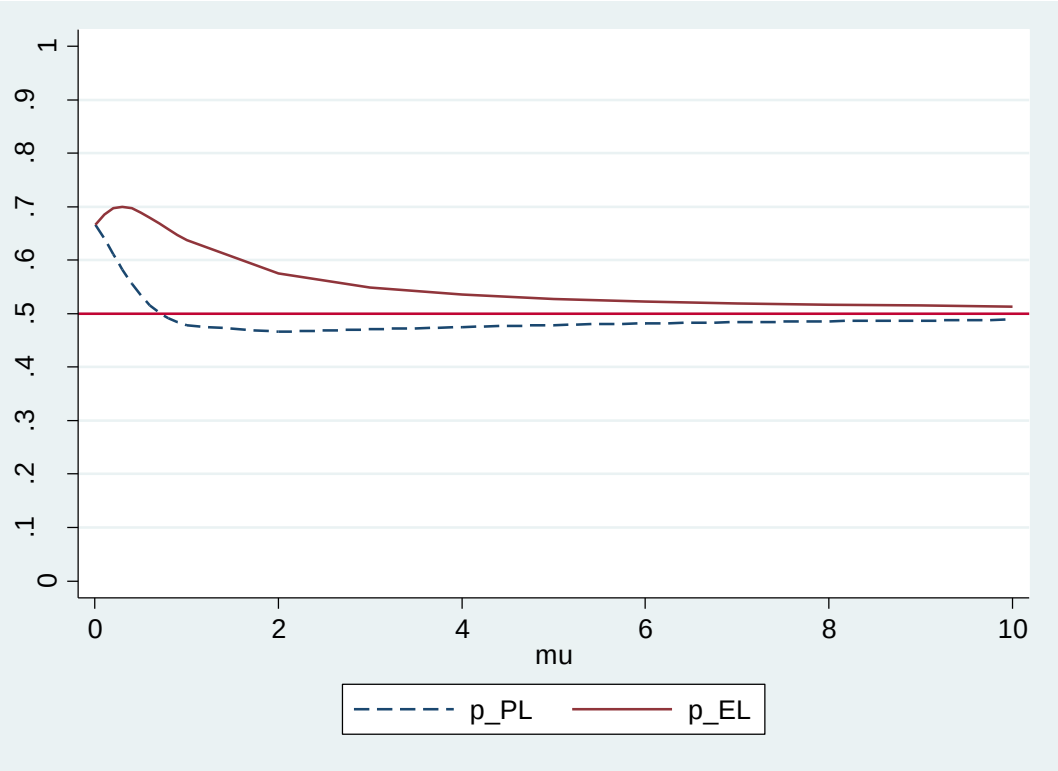

Figure 8.1: Probabilities of Pursuer choosing L and Evader choosing L against "noise level" (μ).

8.5 Estimation of the QRE Model

Having obtained the two probabilities, $p_{PL}(\mu)$ and $p_{EL}(\mu)$, by the procedure described in Section 8.4, it is a simple matter to construct the log-likelihood function. Let $y_{P,it} = 1$ if the proposer in pair i chooses L in round t, and 0 otherwise. Define $y_{E,it}$ likewise for the evader in pair i and round t. Then the log-likelihood function for the complete sample of nT realisations of the game is:

$$LogL(\mu) = \sum_{i=1}^{n}\sum_{t=1}^{T} \ln\left[\left\{y_{P,it}p_{PL}(\mu)+\left(1-y_{P,it}\right)\left(1-p_{PL}(\mu)\right)\right\}\left\{y_{E,it}p_{EL}(\mu)+\left(1-y_{E,it}\right)\left(1-p_{EL}(\mu)\right)\right\}\right] \quad (8.12)$$

If n_{PL} is defined as the number of times, out of the total number of rounds played, n, that Pursuers choose L, and n_{PR}, n_{EL} and n_{ER} are defined similarly, then the log-likelihood function (8.12) may be written much more compactly as:

$$LogL(\mu) = n_{PL}\ln p_{PL}(\mu) + n_{PR}\ln\left[1-p_{PL}(\mu)\right] + n_{EL}\ln p_{EL}(\mu) + n_{ER}\ln\left[1-p_{ER}(\mu)\right] \quad (8.13)$$

The presentation of the log-likelihood function in the form (8.13) makes it clear that the only pieces of information that are required to estimate the parameter μ are the frequencies of the choices; no information is required on the actual sequences of choices. Accordingly, we may say that the choice frequencies are *sufficient statistics* for the estimation of μ.

What is unusual about the estimation process for QRE is that it consists of one optimisation routine inside another. The probabilities used in the construction of the log-likelihood

function (8.13) are themselves computed by the numerical optimisation routine outlined in Section 8.4

9. Models of Learning

The last section was concerned with the modelling of repeated game data, and in particular how to allow for departures from equilibrium. The models considered there were static, in the sense that there was no scope for strategies to change over time. In this Section, we allow for the possibility that players are far from equilibrium in the early rounds of a game, but, in the course of the experiment, adapt or evolve, either part of the way or all of the way, towards the equilibrium. The process which causes this evolution towards the equilibrium is a *learning* process.

Learning is defined as an observed change in behaviour on account of the accumulation of experience. Learning processes have already been modelled in earlier sections. For in the risky choice model of Section 6.6, we considered ways of allowing probability weighting parameters and tremble parameters to change over the course of the experiment. However, in that context, learning was only about the task, not about the behaviour of other players, nor even about the outcomes of previous tasks. This is why, in those situations, the modelling of a learning process simply involved allowing particular parameters to depend in some way on the task number.

The situation in repeated interactive games is very different, because subjects typically receive direct feedback in every round. In each round, they observe both the chosen strategy of the opponent, and the outcome in terms of their and others' pay-offs. Hence they learn directly about the behaviour of others, and also about the type of strategies that are most profitable for themselves. The process of learning about other players is complicated by the fact that other players' behaviour changes as they, too, gain experience. A comprehensive model of learning should therefore incorporate the effects of a players past pay-offs, and also the effects of past choices by other players.

Here we will consider a number of different learning models that are popular in Experimental Economics.

9.1 Directional Learning

Directional learning theory was proposed by Selten and Stoecker (1986). It is relatively simple and highly intuitive. It is built on the idea that subjects adjust their behaviour in each period in response to the outcome of the previous period.

Consider the well known prisoner's dilemma game. It is well known that this game has a dominant solution of non-cooperative behaviour, despite cooperation being to the players' mutual advantage. However, cooperation is often seen when the game is played repeatedly. In Selten and Stoecker's (1986) experment, pairs of players play "supergames" consisting of a sequence of ten plays of the game. What is typically observed is cooperation for the majority of the supergame, but defection by one or both players near the end of the sequence. Of interest is how a subject decides in which round to defect. Clearly, the player who defects first secures a payoff advantage over their opponent. However, the player defecting first may feel that they might have enjoyed a greater total payoff by defecting *later*, since they do not know in which later round their opponent would have defected. Similarly, the player who

does not defect first must realise that they would have needed to defect earlier in order to secure the payoff advantage for themselves. If the two players happen to defect in the same round, it is likely that they would both feel that they should have defected one round earlier in order to secure the advantage.

In Selten and Stoecker's (1986) experment, 35 subjects participated in 25 supergames. Players were randomly rematched between supergames. They model their data using a Markov learning model, with changes in defection period determined by transition probabilities which depend on experience in the previous period. Moffatt (2015) adopts a more direct method for capturing the learning process, by simply performing a linear regression with the change in intended defection period as the dependent variable, and variables representing the outcome of the previous period as axplanatory variables. The results from this regression are:

$$\Delta defect_t = 0.36^{**} before_{t-1} - 0.16^{*} same_{t-1} - 0.61^{**} after_{t-1} \qquad (9.1)$$

where $\Delta defect_t$ is the change in defection period between supergames t-1 and t, and the three explanatory variables are dummy variables indicating respectively whether the player defected before, at the same time, or after, their opponent in the previous supergame. Note that the regression exludes an intercept to allow all three of these variables to be present. Stars indicate significance, and double-stars strong significance.

All three estimates in (9.1) are significantly different from zero. The largest coefficient in absolute value is on the lag of "after", and this indicates that if a player's intended defection period is later than their opponent's, they make their own deviation period (on average) 0.61 periods earlier in the following supergame. The coefficient of "before" indicates that if the player deviates earlier than their opponent, they deviate 0.36 periods later in the following supergame. If the players defect in the same period, they defect 0.16 periods earlier in the following supergame. As expected, this coefficient is the smallest (in magnitude) of the three.

These results are similar to those obtained by Selten and Stoecker (1986), and they strongly support directional learning theory: the decision variable changes in a *direction* predicted by the outcome of the previous round.

9.2 Reinforcement Learning

Reinforcement learning (RL) models (Erev and Roth, 1998) have been used primarily by psychologists. These models are built on the assumption that strategies are "reinforced" by their previous payoffs, and the propensity to choose a strategy depends on the cumulative amount of "reinforcement" attributed to the strategy. Note that players who learn by reinforcement do not pay attention to the decisions that other players have made, only to the realised outcomes of their own decisions.

We continue to consider two-player games. The two players are indexed by i (i=1,2). The game is repeated over several rounds indexed t (t=1,…,T). Assume that, as in the pursue-evade game considered earlier, there are two possible strategies for each player s_i^0 and s_i^1. Let $s_i(t)$ be the strategy chosen by player i in round t, and let $s_{-1}(t)$ be that chosen by the other player. Player i's payoff in round t is given by the scalar-valued function $\pi_i\left(s_i(t), s_{-i}(t)\right)$.

The central feature of the RL model (and other learning models) is a set of variables known as "attractions" which are updated following each round. $A_i^j(t), j = 0,1$ represents Player i's

attraction to strategy j, following round t. Players are likely to have relevant experience before the start of the game, and this experience is represented by the "initial attractions" $A_i^j(0)$ which are parameters to be estimated.

In RL, the updating rule for each attraction variable is:

$$A_i^j(t) = \phi A_i^j(t-1) + I\left(s_i(t) = s_i^j\right)\pi_i\left(s_i^j, s_{-i}(t)\right) \quad i = 1,2;\ j = 0,1;\ t = 1,\cdots,T \qquad (9.2)$$

where I(.) is the indicator function. The important feature of (9.2) is that, due to the presence of the indicator function in the final term, a player's attraction to a strategy can only increase if that strategy is chosen, and the attraction increases by the amount of the payoff received from the chosen strategy. The parameter ϕ appearing in the first term is known as the "recency parameter" and indicates the speed at which past payoffs are forgotten. If ϕ=0, only the payoff from the last round matters; if ϕ=1, payoffs from all previous rounds matter equally.

The choice probabilities in round t are obtained from the attractions in t-1 using the logistic transformation (see Camerer and Ho, 1999):

$$P_i^j(t) = \frac{\exp\left(\lambda A_i^j(t-1)\right)}{\exp\left(\lambda A_i^1(t-1)\right) + \exp\left(\lambda A_i^2(t-1)\right)} \quad i = 1,2;\ j = 0,1;\ t = 1,\cdots,T \qquad (9.3)$$

Note that (9.3) satisfies the basic requirement that (for example) the probability of choosing strategy 0 is monotonically increasing in the attraction to strategy 0, and monotonically decreasing in the attraction to strategy 1. The parameter λ represents the sensitivity to attractions: if λ is a large positive number, attractions are important; if λ=0, attractions are irrelevant and the choice probabities are both 0.5.

In total, there are four free parameters in the RL model: $\phi, \lambda, A_1^0(0), A_2^0(0)$. These four parameters are estimated by ML, where the likelihood contributions are provided by (9.3).

9.3 Belief Learning

Belief learning (BL) models, or belief-based models (Cheung and Friedman, 1997) have been used primarily by decision and game theorists. These models are, in contrast to reinforcement models, built on the assumption that players keep track of their opponents' decisions, and form beliefs about their future play based on this observed history. They then select a best response given these beliefs. Note that belief learners do not pay attention to past successes, only to past behaviour by opponents.

The only difference of BL from RL is the way in which attractions are updated. A simple version of the BL updating rule is:

$$A_i^j(t) = A_i^j(t-1) + \pi_i\left(s_i^j, s_{-i}(t)\right) \quad i = 1,2;\ j = 0,1;\ t = 1,\cdots,T \qquad (9.4)$$

According to (9.4), following each round (t), player i's attraction to strategy j simply increases by the pay-off that was, *or would have been*, received, given the choice $s_{-i}(t)$ made by the other player. Model (9.4) is known as the Cournot learning model.

It is, of course, likely that players take account of more than just the behaviour of the other player in the previous round. According to the "fictitious play model", players take account of the other player's behaviour in *all* previous rounds, with equal importance attached to each previous round. More realistic still is the "weighted fictitious play model", in which the experience of recent rounds carries more weight than that of rounds further back in the past. In this, most general form of BL, there are five free parameters: the initial attractions $A_1^0(0), A_2^0(0)$; an "initial experience" parameter N(0); a "recency parameter" ϕ indicating the weighting of past observations; and the sensitivity parameter λ appearing in the probability formula (9.3).

9.4 The Experience Weighted Attraction (EWA) Model

The Experience Weighted Attraction (EWA) model (Camerer and Ho, 1999) is a model that nests RL and BL. It contains a total of seven free parameters: the same parameters as in RL and BL, and also an additional parameter δ that indicates which of RL and BL is closer to the true model.

Wilcox (2006) reports Monte Carlo evidence of the effect of parameter heterogeneity in EWA. He finds that when certain parameters vary across subjects, but such heterogeneity is neglected in estimation, it results in severe bias in the estimation of the parameter δ in such a way as to favour RL over BL. There is a strong hint here that between-subject heterogeneity should be incorporated into learning models, although this may well result in a problem of over-parameterisation.

One obvious way to introduce heterogeneity is to use a mixture model combining RL and BL. Note that EWA is not a mixture model; it is a representative agent model built on the assumption that all players learn according to the same weighted combination of RL and BL. A mixture model would instead assume that some players learn by RL, and others by BL. This is for future research.

Moffatt (2015) has explained in detail how EWA can be estimated using STATA. Since RL and BL are both nested within EWA, he also demonstrates that both theories can be tested easily using LR tests of restrictions on the EWA parameters.

10. Conclusion

The purpose of this concluding section is to provide a summary of the Monograph, and, more importantly, to attempt to identify emerging themes.

The first part of the survey was concerned with straightforward treatment testing and the closely related topic of power analysis. There is no doubt that power analysis is rapidly taking hold in experimental economics as it appears that many researchers are now treating it as a routine step, both at the design stage, and again in the statistical analysis of the results. One reason for the growing importance of power analysis is its key role in the scientific quality debate, which has recently had a major impact across disciplines. It is possible that power analysis is an unfamiliar concept to mainstream econometricians, since the subject of econometrics has developed largely on the basis of passive data collection. For such researchers, Part 3 of the Monograph may hopefully serve as a useful introduction to the topic.

The power command in STATA is very useful (and other useful packages for power analysis are available, such as G*power). However, a point that was stressed is that such off-the-shelf packages are only useful for in relatively simple settings. In more complex settings the Monte Carlo approach is required for power calculations, and this has been demonstrated in various ways. Some of the findings may be considered surprising, and this is perhaps something that should also become more routine.

Another sense in which basic power analysis might be considered inadequate is that it simply provides a "required sample size", while in many experimental settings, the sample size is at least a two-dimensional concept: number of subjects and number of tasks. Monte Carlo results presented in Section 3 suggest that there are situations in which increasing the number of tasks is more beneficial than increasing the number of subjects. This is a useful result. There are other compelling reasons for using repeated tasks. It provides an opportunity for subjects to learn, and it is likely (although by no means obvious) that the researcher is more interested in the behaviour of experienced than inexperienced subjects. However, other behavioural considerations are relevant. For example, the number of tasks should not be increased without considering how many tasks the typical subject may be reasonably required to perform before losing concentration or losing interest.

There are other examples in which behavioural considerations appear to be pulling in the opposite direction from theoretical analysis. An obvious one is the choice between a between-sample and within-sample test. Within-sample tests are clearly more efficient, but researchers prefer between-sample tests because of order effects. Perhaps more work is required assessing how important order effects really are.

Another straighforward example where the most efficient design is not necessarily the best design arises in the analysis of risk attitude. From a theoretical perspective, the most efficient way of eliciting risk attitude is to ask for a certainty equivalent, but for various behavioural reasons this method of elicitation is thought to distort true risk attitudes, and choice between lotteries is the preferred elicitation method. One obvious reason for this is that, for many subjects, lottery valuation is likely to be an unfamiliar task, while the choice between lotteries is a task that resembles many everyday situations.

Yet another example which has been discussed is a situation in which ordinal data on strength of preference has been elicited. Clearly – from a theoretical perspective - ordinal data with many categories leads to more efficient estimation than binary data. But, as discussed, it is hard, or impossible, to incentivise ordinal responses, and the question therefore arises of the validity of such responses.

The point that is emerging here is that in the design of behavioural experiments, there is perhaps a broader range of considerations than for scientific experiments.

We also considered optimal designs for binary choice experiments. This amounts to the design of the choice problems themselves, e.g. the probabilities defining the lotteries. A problem here, and a problem for experimental design in the context of nonlinear models generally, is that the parameters of interest need to be known in order to constuct the optimal choice problems (the well-known "chicken-and-egg" problem). This problem is analogous to the problem of needing to know the true effect size in order to find the required sample size for a treatment test. The message is that researchers need to look for priors, from previous research whenever available, to inform design decisions.

Another theme has been decisions of stochastic specification. Choice of stochastic specification is particularly important in the modelling of risky choice decisions. The Fechner Model seems to be a useful approach, but the non-monotonicity problem outlined in Section 6.9 appears at first sight to be quite damning. It is possible that this problem is avoided by altering the stochastic specification and this suggests an important area for research. The Random Preference model avoids the non-monotonicity problem, but also as discussed in Section 6.9, it is hard to apply it to complex designs.

Yet another theme has been the modelling of heterogeneity. When the decision variable has a continuous distribution, such as bids in auctions or effort in contests, the multilevel model provides a relatively straightforward, and reliable, means of allowing for the between-subject, between-group, and between-session heterogeneity that is a feature of so many such experiments. In situations in which the decision variable has more complicated features (e.g. choices between alternatives), the method of maximum simulated likelihood has been proposed as a suitable method for dealing with heterogeneity, and one that is particularly useful when there is more than one dimension of heterogeneity (e.g. risk aversion and probability weighting). The finite mixture framework has been demonstrated as a means of separating subjects into discrete types.

One theme that was only touched upon is that of the *process* of decision making. Thoughout most of the monograph we have been concerned only with the decisions made by experimental subjects. There is a growing interest in Experimental Economics in the process leading to the decision. One obvious means of capturing process is to analyse decision times (Moffatt, 2005; Alós-Ferrer et al., 2016). But other aspects of process such as the use of eye-tracking to predict choices (e.g. Jiang et al., 2016) are rapidly gaining popularity.

Appendix A. MPL Lottery Choice Problems from Conte et al. (2019)

	Lottery A				Lottery B				Please Circle EITHER A or B Stop when First B Circled	
Row	probability ($400)	Payoff	probability ($100)	Payoff	probability ($?)	Payoff	probability ($50)	Payoff	Choose A	Choose B
1	0.3	$400	0.7	$100	0.1	$680	0.9	$50	A	B
2	0.3	$400	0.7	$100	0.1	$750	0.9	$50	A	B
3	0.3	$400	0.7	$100	0.1	$830	0.9	$50	A	B
4	0.3	$400	0.7	$100	0.1	$930	0.9	$50	A	B
5	0.3	$400	0.7	$100	0.1	$1,060	0.9	$50	A	B
6	0.3	$400	0.7	$100	0.1	$1,250	0.9	$50	A	B
7	0.3	$400	0.7	$100	0.1	$1,500	0.9	$50	A	B
8	0.3	$400	0.7	$100	0.1	$1,850	0.9	$50	A	B
9	0.3	$400	0.7	$100	0.1	$2,200	0.9	$50	A	B
10	0.3	$400	0.7	$100	0.1	$3,000	0.9	$50	A	B
11	0.3	$400	0.7	$100	0.1	$4,000	0.9	$50	A	B
12	0.3	$400	0.7	$100	0.1	$6,000	0.9	$50	A	B
13	0.3	$400	0.7	$100	0.1	$10,000	0.9	$50	A	B
14	0.3	$400	0.7	$100	0.1	$17,000	0.9	$50	A	B

	Lottery A				Lottery B				Please Circle EITHER A or B Stop when First B Circled	
Row	probability ($400)	Payoff	probability ($300)	Payoff	probability ($?)	Payoff	probability ($50)	Payoff	Choose A	Choose B
1	0.9	$400	0.1	$300	0.7	$540	0.3	$50	A	B
2	0.9	$400	0.1	$300	0.7	$560	0.3	$50	A	B
3	0.9	$400	0.1	$300	0.7	$580	0.3	$50	A	B
4	0.9	$400	0.1	$300	0.7	$600	0.3	$50	A	B
5	0.9	$400	0.1	$300	0.7	$620	0.3	$50	A	B
6	0.9	$400	0.1	$300	0.7	$650	0.3	$50	A	B
7	0.9	$400	0.1	$300	0.7	$680	0.3	$50	A	B
8	0.9	$400	0.1	$300	0.7	$720	0.3	$50	A	B
9	0.9	$400	0.1	$300	0.7	$770	0.3	$50	A	B
10	0.9	$400	0.1	$300	0.7	$830	0.3	$50	A	B
11	0.9	$400	0.1	$300	0.7	$900	0.3	$50	A	B
12	0.9	$400	0.1	$300	0.7	$1,000	0.3	$50	A	B
13	0.9	$400	0.1	$300	0.7	$1,100	0.3	$50	A	B
14	0.9	$400	0.1	$300	0.7	$1,300	0.3	$50	A	B

References

Alós-Ferrer, C., Granić, Đ. G., Kern, J., & Wagner, A. K. (2016). Preference reversals: Time and again. *Journal of Risk and Uncertainty*, *52*(1), 65-97.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, *110*(1), 1-21.

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293-312.

Arrow, K. J., & Enthoven, A. C. (1961). Quasi-concave programming. *Econometrica: Journal of the Econometric Society*, 779-800.

Atkinson, A. C. (1996). The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 59-76.

Allais, M. (1953). "Le Comportement de l 'Homme Rationnel devant le Risque: Critique des Postulats et axiomes de l'ecole Americaine", *Econometrica* 21, 503-546.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, *9*(4), 383-405.

Apesteguia, J., & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74-106.

Bacon, P. M., Conte, A., & Moffatt, P. G. (2020). A test of risk vulnerability in the wider population. *Theory and Decision*, *88*(1), 37-50.

Bardsley, N. (2008). Dictator game giving: altruism or artefact?. *Experimental Economics*, *11*(2), 122-133.

Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.

Bardsley, N., & Moffatt, P. G. (2007). The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, *62*(2), 161-193.

Bardsley N. and Moffatt P.G. (2020). A Meta Analysis of the Preference Reversal Phenomenon. WP, University of East Anglia.

Ben-Ner, A., Kong, F., & Putterman, L. (2004). Share and share alike? Intelligence, socialization, personality, and gender-pairing as determinants of giving. *Journal of Economic Psychology*, *25*(5), 581-589.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Systems Research and Behavioral Science*, *9*(3), 226-232.

Berenson M.L., Levine D., Rindskopf, D. (1988), *Applied Statistics: A first course*, Prentice Hall, New York.

Blavatskyy, P. R. (2011). Probabilistic risk aversion with an arbitrary outcome set. *Economics Letters*, *112*(1), 34-37.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review*, *90*(1), 166-193.

Bosch-Domènech, A., Montalvo, J. G., Nagel, R., & Satorra, A. (2010). A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental economics*, *13*(4), 461-475.

Bosman, R., & Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *The Economic Journal*, *112*(476), 147-169.

Butler, D., Isoni, A., Loomes, G., & Navarro-Martinez, D. (2014). On the measurement of strength of preference in units of money. *Economic Record*, *90*(s1), 1-15.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827-874.

Camerer, C., Ho, T., & Chong, K. (2003). Models of thinking, learning, and teaching in games. *The American Economic Review*, *93*(2), 192-195.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861-898.

Camerer, C. F., et al. (2016) "Evaluating replicability of laboratory experiments in economics." *Science* 351.6280: 1433-1436.

Camerer, C. F., Dreber, A., & Johannesson, M. (2019). Replication and other practices for improving scientific quality in experimental economics. In *Handbook of Research Methods and Applications in Experimental Economics*. Edward Elgar Publishing.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818-827.

Cheung, Y. W., & Friedman, D. (1998). A comparison of learning and replicator dynamics using experimental data. *Journal of economic behavior & organization*, *35*(3), 263-280.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218-1221.

Chong, J. K., Ho, T. H., & Camerer, C. (2016). A generalized cognitive hierarchy model of games. *Games and Economic Behavior*, *99*, 257-274.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, *47*(1), 72-93.

Cohen, J. (2013), *Statistical power analysis for the behavioural sciences*, Routledge Academic, London.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Conlisk, J. (1989). Three variants on the Allais example. *The American Economic Review*, 392-407.

Connolly, T., & Butler, D. (2006). Regret in economic and psychological theories of choice. *Journal of Behavioral Decision Making*, *19*(2), 139-154.

Conte, A., Hey, J. D., & Moffatt, P. G. (2011). Mixture models of choice under risk. *Journal of Econometrics*, *162*(1), 79-88.

Conte, A., Moffatt, P. G. (2014). The econometric modelling of social preferences. *Theory and decision*, *76*(1), 119-145.

Conte, A., Levati, M. V., & Nardi, C. (2018). Risk preferences and the role of emotions. *Economica*, *85*(338), 305-328.

Conte A., Moffatt P.G., Riddel M. (2019), "The Multivariate Random Preference Estimator for Switching Multiple Price List Data", School of Economics Discussion Paper 19-04, University of East Anglia.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 829-844.

Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, *19*(3), 613-641.

Daniels, R. C., & Rospabé, S. (2005). Estimating an earnings function from coarsened data by an interval censored regression procedure. *Studies in Economics and Econometrics*, *29*(1), 29-45.

Daykin, A. R., & Moffatt, P. G. (2002). Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(3), 157-166.

Eckel, C. C., & Grossman, P. J. (1998). Are women less selfish than men?: Evidence from dictator experiments. *The economic journal*, *108*(448), 726-735.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall, New York.

El-Gamal, M. A., & Grether, D. M. (1995). Are people Bayesian? Uncovering behavioral strategies. *Journal of the American statistical Association*, *90*(432), 1137-1145.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583-610.

Engel, C. (2020). Estimating heterogeneous reactions to experimental treatments. *MPI Collective Goods Discussion Paper*, (2019/1).

Engel, C., & Moffatt, P. G. (2012). Estimation of the house money effect using hurdle models. *MPI Collective Goods Preprint*, (2012/13).

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, *94*(4), 857-869.

Epps, T. W., & Singleton, K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, *26*(3-4), 177-203.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848-881.

Fallucchi, F., Luccasen, R. A., & Turocy, T. L. (2019). Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering. *Journal of the Economic Science Association*, *5*(2), 238-254.

Fallucchi, F., Mercatanti, A., & Niederreiter, J. (2020). Identifying types in contest experiments. *International Journal of Game Theory*, 1-23.

Fechner, G., (1860). *Elements of Psychophysics, Vol 1*, Holt, Rinehart and Winston, New York.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817-868.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, *6*(3), 347-369.

Fréchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics*, *15*(3), 485-498.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651.

Gächter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, *13*(3), 364-377.

Georg S.J. (2009). "Nonparametric testing of distributions – the Epps-Singleton two-sample test using the empirical characteristic function", *The Stata Journal*, 9, 454-465.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The german socio-economic panel (soep). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345-360.

Greene W. (2011). *Econometric Analysis*. 7th Edition. Pearson, London.

Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Goeree, J. K., Holt, C. A., & Palfrey, T. R. (2003). Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, *45*(1), 97-113.

Gollier, C., & Pratt, J. W. (1996). Risk vulnerability and the tempering effect of background risk. *Econometrica: Journal of the Econometric Society*, 1109-1123.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38(1):129–166.

Gourieroux, C., Monfort, A., (1996). *Simulation-based econometric methods*. Oxford university press.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, *69*(4), 623-638.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, *2*(1), 84-90.

Hammersley, J. M.; Handscomb, D. C. (1964). *Monte Carlo Methods.* Methuen.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2005). Risk aversion and incentive effects: Comment. *American Economic Review*, 897-901.

Harrison, G. W., & Rutström, E. E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental economics*, *12*(2), 133.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, 1291-1326.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251-1271.

Holt, C. A., & Laury, S. (2002). "Risk aversion and incentive effects". *American Economic Review*, 92, 1644-1655.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *The American economic review*, *95*(3), 902-904.

Huber, J., & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing research*, *33*(3), 307-317.

Ioannidis J.P.A., Stanley T.D., Doucouliagos H. (2018). The power of bias in economics research. The *Economic Journal* 127: F236-265.

Isoni, A., Loomes, G., & Sugden, R. (2011). The willingness to pay—willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Comment. *The American Economic Review*, *101*(2), 991-1011.

Jakiela, P. (2013). Equity vs. efficiency vs. self-interest: on the use of dictator games to measure distributional preferences. *Experimental Economics*, *16*(2), 208-221.

Jiang, T., Potters, J., & Funaki, Y. (2016). Eye-tracking social preferences. *Journal of Behavioral Decision Making*, *29*(2-3), 157-168.

Johnson C., Baillon, A., Li, Z., van Dolder, D., & Wakker, P. P. (2019). Prince: An Improved Method for Measuring Incentivized Preferences.

Kahneman D. and Tversky A. (1979) Prospect Theory: an analysis of Decisions under Risk. *Econometrica* 47(2), 263-291.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of political Economy*, *98*(6), 1325-1348.

Keasey, K., & Moon, P. (1996). Gambling with the house money in capital expenditure decisions: An experimental analysis. *Economics Letters*, *50*(1), 105-110.

Keynes J.M. (1936). The General Theory of Employment Interest and Money. Macmillan.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299-312.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482-493.

List, J., Sadoff, S., Wagner, M., 2011. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Experimental Economics 14, 439-457

Loomes, G., & Sugden, R. (1998). Testing different stochastic specificationsof risky choice. *Economica*, *65*(260), 581-598.

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of risk and Uncertainty*, *24*(2), 103-130.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, *66*(2), 81.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, *35*(4), 615-645.

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Cambridge university press.

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, *10*(1), 6-38.

Moffatt, P. G. (2007). Optimal Experimental Design in Models of Decision and Choice. In Boumans M. (ed.) *Measurment in Economics: A Handbook*, Academic Press: London, 357-375.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for experimental economics*. Palgrave Macmillan.

Moffatt P.G. (2020) The Experimetrics of Depth of Reasoning Models, in *Handbook of Experimental Game Theory.* Capra, M., Croson, R., Rigdon, M. & Rosenblat, T. (eds.). Edward Elgar.

Moffatt, P. G., & Peters, S. A. (2001). Testing for the presence of a tremble in economic experiments. *Experimental Economics*, *4*(3), 221-228.

Moffatt, P. G., & Zevallos-Porles, G. (2020). A Kuhn-Tucker Model for Behaviour in Dictator Games. Discussion Paper (No. 20-03). School of Economics, University of East Anglia, Norwich, UK.

Moffatt, P.G., Pogrebna G., and Zevallos-Porles G. (2020). Depth of Reasoning Models with Sophisticated Agents. Discussion Paper (No. 20-XX). School of Economics, University of East Anglia, Norwich, UK..

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, *32*(3), 385-397.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313-1326.

Nagel, R., Bühren, C., & Frank, B. (2017). Inspired and inspiring: Hervé Moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*, *90*, 191-207.

Nelson F.D. (1976). On a general computer algorithm for the analysis of models with limited dependent variables. *Annals of Economic and Social Measurement*, 5, 493-509.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27-29.

Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, *22*(4), 1002-1029.

Plott C.R., Zeiler, K. (2007). Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory?. *The American Economic Review*, *97*(4), 1449-1466.

Pratt J. (1964). Risk Aversion in the Small and in the Large. *Econometrica* 32, 122-136.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323-343.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Rosenthal, R. W., Shachat, J., & Walker, M. (2003). Hide and seek in Arizona. *International Journal of Game Theory*, *32*(2), 273-293.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068-1095.

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method* (Vol. 10). John Wiley & Sons.

Runco, M. (2013). Estimating depth of reasoning in a repeated guessing game with no feedback. *Experimental Economics*, *16*(3), 402-413.

Saha, A. (1993). Expo-power utility: a 'flexible'form for absolute and relative risk aversion. *American Journal of Agricultural Economics*, *75*(4), 905-913.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, *2*(6), 110-114.

Shachat, J., Swarthout, J. T., & Wei, L. (2015). A hidden Markov model for the detection of pure and mixed strategy play in games. *Econometric Theory*, *31*(4), 729-752.

Siegel S., Castellan N.J., (1988), *Non-parametric statistics for the behavioral sciences*, Second edition, McGraw Hill, New York.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, *72*(5), 923-955.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218-254.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, *38*(2), 332-382.

StataCorp. 2019. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

Stewart, M. B. (1983). On least squares estimation when the dependent variable is grouped. *The Review of Economic Studies*, *50*(4), 737-753.

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and uncertainty*, *32*(2), 101-130.

Su, L., Shi, Z., & Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, *84*(6), 2215-2264.

Sun, T., & Ding, Y. (2019). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*.

Tanaka, T., C.F. Camerer, and Q. Nguyen. 2010. Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam, *American Economic Review,* 100, 557-71.

Thaler, R. H., & Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, *36*(6), 643-660.

Train Kenneth, E. (2003). Discrete choice methods with simulation. *Cambridge: Cambridge UniversityPress*.

Tversky, A. and Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102:269–283.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review*, 204-217.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297-323.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.

Wilcox, N. T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, *74*(5), 1271-1292.

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In *Risk aversion in experiments* (pp. 197-292). Emerald Group Publishing Limited.

Wilcox, N. T. (2011). 'Stochastically more risk averse:'A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89-104.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217-235.

Zhang, L., & Ortmann, A. (2013). Exploring the meaning of significance in experimental economics. *UNSW Australian School of Business Research Paper*, (2013-32).