# Chapter 1
# Introduction and Overview

## 1.1 What is Experimetrics?

*Experimetrics*[1] comprises the body of econometric techniques that are customized to experimental applications. A wide variety of such techniques appear in the experimental economics literature. The aim of this textbook is to assemble this body of techniques, to demonstrate their use in a hands-on style, drawing on as wide a range of examples as possible, and to interpret each set of results in ways that are most useful to experimental economists. The target audience is mainly researchers in experimental economics. It is also conceivable that the book might be of interest to econometricians who are curious to know what sort of techniques are being used by experimental economists.

The experimetric techniques that already appear in the experimental economics literature range from the very basic to the highly sophisticated. At the basic end of this spectrum we see the class of techniques known as *treatment tests*, that is, tests which compare outcomes with and without a treatment, or before and after a treatment. At the sophisticated end of the spectrum, we see highly complex structural models, with a deterministic core corresponding to an underlying behavioural theory, with possibly many structural parameters, and a stochastic specification including possibly many dimensions of variation, both within and between subject. Needless to say, the type of econometric approach that is chosen is often, and justifiably, dictated by the type of experiment that has been conducted, and by the types of research questions being addressed.

In some experiments, the "home-grown" characteristics of the experimental subjects are the focus, and the objectives are usually simply to investigate how individuals make decisions, or interact with each other, in particular settings. These

[1] The word "Experimetrics" was (to the best of my knowledge) coined by Camerer (2003, p. 42). Houser's (2008) entry in the *New Palgrave Dictionary of Economics* on "Experiments and Econometrics" commences with the line "'Experimetrics' refers to formal procedures used in designed investigations of economic hypotheses." Bardsley & Moffatt (2007) are apparently the first authors to have used the word in the title of a published paper.

studies typically rely on relatively simple experimental designs (e.g. choosing between lotteries; splitting a pie), with the ultimate objective of measuring subjects' characteristics, especially preference parameters. It is normal to expect substantial variation in these measured characteristics. Indeed, it is often the precise features of this variation in which we are most interested; for example, the proportion of the population who are "selfish", or the proportion who are EU maximisers. When data are from this type of experiment, it is often seen as appropriate to model the decision-making process using structural estimation methods, for example methods that simultaneously estimate all of the parameters appearing in the individuals' objective function, as well as distributional parameters, some of which capture preference heterogeneity.

In other experiments, the focus is on the functioning of an economic institution, rather than the characteristics of the individual participants within it, and the objective may be to test a particular theory as applying to that institution. In these settings, *induced value methodology* is commonly used. This technique is based on the idea that the appropriate use of a reward medium allows the experimenter to *induce* pre-specified characteristics (e.g. preferences) in the subjects so that their innate characteristics become irrelevant. Having essentially eliminated the impact of subjects' characteristics in this way, it is clearly much easier to apply close scrutiny to the theory under test. In these settings, the experimental designs are relatively complex, since key features of the economic institutions need to be captured in convincing ways. However, the econometric techniques required are often very simple. The level of control is normally such that straightforward treatment testing is often seen to be the natural framework, and the most suitable means of obtaining answers to the research questions of interest.

The following subsections provide brief overviews of the types of econometric techniques that are best suited to each of these two broad areas. As such, this chapter provides a broad overview of the contents of the remainder of the book.

## 1.2 Experimental Design

The topic of experimental design is clearly much more important in experimental economics than in other areas of economics. This is because, in other areas, typically the data generation process is out of the control of the investigator. In experimental economics, the data generation process is very much within the control of the investigator. Hence, design issues such as the choice of sample size, the sampling process, and the process for assignment of subjects to treatments, all take centre stage.

A central concept is randomization. If randomization is correctly applied in the process of selecting subjects for an experiment, then identification of treatment effects, which has always been a central problem in mainstream econometrics, is not a problem. The other side of this coin is that, since the data have not been collected in a natural environment, experimental results do not necessarily carry over to the world outside the lab. We therefore see that the advantages of experimental research

in terms of identification may be seen to be countered by the disadvantage of non-generalisability (see Al-Ubaydli & List, 2013).

There are several different types of design, including completely randomised designs, within-subject designs, crossover designs, and factorial designs. When groups of subjects are playing against each other, a choice also needs to be made between "partner" and "stranger" matching. Each design has both advantages and disadvantages, and the decision of which to use is often a delicate issue.

The choice of sample size is another key design decision. The question to be addressed here is how many subjects are required for the experimenter to be confident of reliable conclusions. More precisely, how many are required in each treatment? A useful framework for addressing these questions is power analysis (Cohen, 2013); that is, using probability theory to find what sample is required to provide a given "power" of the test being conducted.

A particularly interesting problem in experimental design is how to specify binary choice problems (e.g. lottery choice) in such a way as to generate a data set from which subjects' preference parameters may be estimated with maximal precision. A chapter of the text is dedicated to this problem.

## 1.3 The Experimetrics of Theory Testing

It is often claimed the purpose of an experiment is to test an economic theory. From the econometrician's perspective, a natural way to perceive such a test is as an assessment of whether the *predictions* from the theory provide good approximations to actual behaviour (i.e. the behaviour of subjects in the lab). From this perspective, the role of the experimenter is to find regularities in observed behaviour, and then to ask which theories are best able to account for these regularities.

Competitive equilibrium is the central concept in many theories. The objective of an experiment (a market experiment, say) may be simply to observe how close behaviour is to the competitive equilibrium, and we shall refer to this as testing the *fundamental prediction* of the theory. This type of experiment is usually performed using *induced value* methodology; that is, a system in which each subject is exogenously assigned a valuation of the object being traded, so that the complete demand and supply schedules, and therefore the equilibrium, are known by the experimenter. In this setting it is clearly a simple matter to use experimental data to assess how close actual behaviour is to the fundamental (equilibrium) prediction.

An important aspect of the fundamental prediction of the theory is that it often amounts to a "point prediction", that is, it simply tells us that a particular decision variable will, given the known fixed values of the exogenous variables, take a particular value. There are no *free parameters* in the model that generates the fundamental prediction. To see how free parameters enter a model, consider the following example. Start with a benchmark model that is built on the assumption of risk neutrality (RN). Such a model is likely to lead to a "risk-neutral equilibrium" prediction, $y = constant$, where $y$ is the decision variable. This benchmark model has no free parameters. Now consider what happens when we adopt the assumption of expected

utility (EU) maximisation in place of the RN assumption. This inevitably results in the appearance of (at least) one free parameter, which will typically be one of the standard measures of risk aversion. The model's prediction will now depend on the value taken by this free parameter. Next, consider what happens if we futher generalise the model to assume that subjects behave in accordance with prospect theory (Kahneman & Tversky, 1979) instead of EU. This extension results in the further addition of free parameters capturing probability weighting and loss aversion.

Fairly early in the book, we shall be drawing heavily on the examples of experimental auctions and contests. Auction theory and contest theory are both well developed and lead to very clear predictions. The fundamental prediction in these contexts usually takes the form of risk-neutral Nash equilibrium (RNNE) bidding behaviour, which depends on the precise structure of the type of auction or contest under study. With experimental data on subjects' bids, it is usually a simple matter to test whether behaviour is consistent with the RNNE.

In many situations, behaviour of experimental subjects tends to depart in systematic ways from the "fundamental prediction" of theory. In the context of auctions and contests, these departures take the form of systematic "over-bidding" relative to the Nash equilibrium prediction. Hence, if our only objective were to test the fundamental predictions of Nash equilibrium theories, this objective would be straightforwardly met: we would reject the theory. However, there are other levels at which theory may be tested. A typical theory gives rise to a number of "comparative static predictions". These are predictions of the decision variable moving in a particular direction in response to an exogenous change in another variable. For example, in many auction contexts, it is predicted that a rise in the number of bidders in the auction has a negative effect on bids. If two sets of experimental auctions are run, one with four bidders, and the other with six, we might test to see whether bids are lower in the auctions with six bidders. If they are, it would be reasonable to conclude that the experimental data is consistent with this particular comparative static prediction of the theory, even if the theory's fundamental prediction fails.

The test just described, for testing the effect of the number of bidders on bids, is an example of a treatment test. Auctions with the lower number of bidders may be referred to as the "control", and those with the higher number of bidders as the "treatment". There are a large number of possible ways of conducting a treatment test, both in terms of experimental design and in terms of the statistical procedure used to compute the test statistic. The two groups of subjects could be separate, in which case it is a between-sample test. Alternatively, subjects could be subjected to both treatments, in which case it is a within-sample test. The approprate choice of statistical test depends on which of these sampling approaches has been followed, as well as on a number of other design features such as the sample size, and the process by which subjects have been divided into sessions and groups.

There is another important use of treatment tests. In situations in which behaviour is found to depart from the fundamental prediction of the theory, it is of obvious interest to know the reason for this departure; these reasons are sometimes referred to as the "behavioral drivers of out-of-equilibrium play". It might be suggested, for example, that the reason for over-bidding in a contest is a "joy of winning". To test this, a treatment would be designed in which the joy of winning is

absent. If bids are lower under this treatment, we may conclude that joy of winning is indeed a cause of over-bidding. Other suggested reasons for over-bidding include other-regarding preferences, risk aversion, and probability distortion.

When a treatment test is performed, the approach can be non-parametric or parametric. Non-parametric tests are sometimes preferred because their validity relies on fewer assumptions. Parametric tests often rely, to some extent, on assumptions such as normality of the underlying data.

Parametric treatment tests may be performed in the context of a linear regression, in which one of the explanatory variables (or the only explanatory variable) is a "treatment dummy". The advantages of this approach are: the coefficient of the treatment dummy is directly interpretable as the treatment effect; the associated t-statistic is the test statistic for the treatment; more than one treatment may be tested at the same time; the effects of other determinants of the outcome may be controlled for; and regression-related routines, such as clustering of standard errors, may be exploited.

## 1.4 Dependence in Experimental Data

Dependence is an issue that is central to Experimetrics. Basic treatment tests (and many other testing and estimation procedures) rely crucially on the assumption of independent observations. There are several reasons why this assumption can be expected to fail when analysing experimental data (for a recent discussion, see Fréchette, 2012). Firstly, if, as is common, experimental subjects are engaging in a sequence of tasks, there is likely to be "clustering" at the subject level, since some subjects will simply be predisposed to higher values of the decision variable than other subjects. This will of course mean that there are positive correlations between the observations for a given subject. There is also likely to be clustering at the level of the "group" in which a subject operates; a subject's behaviour is likely to depend on the behaviour of other members in the same group. It is often suggested that there is also clustering at the session level; for example, it may be expected that behaviour in the afternoon session differs from that in the morning session, purely as a consequence of the time of day. A more subtle reason for clustering at the session level is when "stranger" treatments are used (Andreoni, 1988), where group composition changes between rounds. Subject-level clustering is the "lowest level" of clustering, while session-level clustering is the "highest level".

There are a number of strategies that may be followed to allow for clustering, in order to validate the tests being performed. One is an ultra-conservative approach: to take the average behaviour over independent units (subjects, groups, or sessions, depending on the level of clustering assumed), and then apply the testing procedure to these averages. Provided the level of clustering at which averaging is performed is sufficiently high, the averages will automatically meet the requirements of independence. The most conservative possible approach would be to take averages at the highest level of clustering. The obvious disadvantage of this approach is that the process of averaging severely reduces the size of the sample to which the test may

be applied, and therefore reduces the test's power (i.e. the probability of finding a significant treatment effect when a treatment effect indeed exists).

The second possible approach is to run regressions with treatment dummies, and to use cluster-standard errors (that is, standard errors corrected for clustering at the assumed level) in the computation of the treatment test statistic. The third possible approach is to use panel data estimation techniques, such as random effects and fixed effects models. These techniques deal directly with the panel structure of the data. It is possible to go even further, by using the multi-level modelling approach, which extends the random effects framework to situations in which there is more than one level of dependence (subject-level and session-level, for example).

## 1.5 Parametric versus Non-parametric Approaches

A fundamental choice is that between parametric and non-parametric methods. One of the key issues in this choice is the scale of measurement of the outcome variable (nominal, ordinal, or cardinal). Many parametric tests rely on distributional assumptions which can only hold if the variables under analysis are measured on a cardinal scale.

Most commonly, the distributional assumption that is required is that of normality in the outcome variable. If the outcome variable is normally distributed, this is highly convenient because it means that, subject to certain other requirements, parametric tests, for example the t-test, can be relied upon. However, data from economic experiments are often clearly non-normal, and this appears to be of particular concern to many experimental economists. Many researchers unquestioningly apply non-parametric tests to their data and explain this choice by the concern that their outcome variable is non-normal. However, this strategy is likely to be costly. When non-parametric tests are applied to cardinal data, the cardinal information in the data is disregarded, since the tests are based solely on the ordinality of the data. This is one reason why non-parametric tests tend to be less powerful (i.e. having a lower probability of detecting an effect when one exists) than their parametric counterparts.

There is a definite sense in which the normality requirement is taken too seriously by experimental economists. Even if the data are non-normal, provided that the sample size is sufficiently large (usually taken to mean more than 30 observations in each treatment), the central limit theorem may be invoked (implying that the *standardized mean* of the data is normally distributed in repeated samples) and parametric tests may be relied upon. Moreover, even if the sample size is insufficient for the central limit theorem to apply, methods are available which enable the valid use of parametric tests. One such method is the bootstrap (Efron & Tibshirani, 1993). This method provides a means of validly conducting parametric tests that respect cardinality, without making any assumptions about the distribution of the data. Hence the bootstrap may be viewed as a means of combining the benefits of parametric and non-parametric tests, whilst avoiding the drawbacks.

Another type of non-parametric method is the non-parametric regression. This is a procedure which, roughly speaking, results in a smooth, flexible curve being

fitted through a scatter plot. It is very useful at the initial, exploratory, stages of data analysis, and is used to determine the nature of the relationship between two variables. In particular, it can be used to determine whether the relationship is linear, and if not, whether it is U-shaped, inverted-U-shaped, cubic, and so on. Finding out the nature of the relationship in this way is clearly very useful in deciding on an appropriate specification for the parametric model. The non-parametric regression technique adopted in this book is the locally-weighted regression (Lowess) technique of Cleveland (1979), which is available in STATA.

## 1.6 Structural Experimetrics

One of the principal objectives of this textbook is to encourage the wider use of fully structural models and to present complete explanations of how they can be estimated.

A structural econometric model is a model that combines an explicit economic theory with an appropriate statistical model. See Reiss & Wolak (2007) for a thorough survey of the applications of structural models in the area of industrial organisation. In the context of Experimetrics, structural modelling becomes potentially very useful when the objective of the experiment is the measurement of "home-grown" features of experimental subjects. By this we mean features such as risk-attitude or degree of altruism. Many would agree that these features are best modelled from the starting point of a utility function: a von Neumann-Morgenstern utility function if risk attitude is the focus; a "two-good" utility function over own payoff and other's payoff if altruism is the focus.

One obvious advantage of the structural approach is it provides a solution to the dependence problem, similar to that provided by multi-level modelling. Under the structural approach, the hypotheses of interest may be validly tested using data at the level of individual decisions. Subject characteristics may be used as explanatory variables to explain "observed" subject-heterogeneity, while a random effect term may be used to account for unobserved heterogeneity. Group or session level random effects may also be included. The structural approach actually allows alternative strategies for controlling for behaviour within the group, for example the inclusion of mean of group contribution in previous round as an explanatory variable.

The scale of measurement is important in the choice between modelling strategies. Much of mainstream econometric modelling is built on the assumption that the outcome is continuous. In experimental economics, this is frequenty not the case. Often the outcome is binary. Sometimes, it is discrete but with more than two outcomes, in which case, we need to consider whether the outcome is nominal (i.e. categorical) or ordinal. Sometimes the outcome from the theoretical model is a continuously distributed variable, but the nature of the data is such that the observed variable is far from continuous: it may be lower or upper censored, or there may be accumulations of data at particular interior "focal points". Even in a situation in which the outcome is truly continuous, we often need to pay careful attention to its distribution. In a fully parametric structural model, it is important

to specify all distributional assumptions correctly in order to achieve consistency of parameter estimates.

In some situations, the structural parameters of interest may be obtained using estimation routines that are readily available in software packages. Such routines include linear regression, panel models, binary data models, censored data models, interval regression, and ordinal models. In other situations, the required routines are not readily available, and purpose-built programs need to be written. Fortunately, software facilitating the development of such programs is readily available. One such tool that is used particularly heavily in this book is the `ml` routine in STATA (Gould et al., 2010).

A considerable advantage of structural models is that they incorporate randomness in behaviour in ways that are natural. Stochastic terms, or "error terms", are something that some experimental economists begrudgingly perceive as a component that needs to be appended to their cherished economic theory, in order for it to be able to explain actual behaviour. Moreover, there is a tendency for them to perceive deviations from expected behaviour as (genuine) "errors" (i.e. mistakes). By encouraging experimental economists to embrace the framework of structural models, we will enable them to move forwards in accepting that randomness is a perennial (and natural) feature of human behaviour. The key message here is that the stochastic term is not an afterthought that needs to be appended; it is an integral feature of the model. To quote Harrison et al. (2015), "in short, one cannot divorce the job of the theorist from the job of the econometrician".

Once the importance of the role of the stochastic specification has been established, the next issue to be addressed is where and how the stochastic elements should be introduced. As we shall see, there are a number of possible approaches. In the context of individual decision making, the most obvious approach is simply to apply an additive error to the equation that represents the theoretical prediction; such an approach is analogous to straightforward regression analysis, although it is sometimes much more complicated than this. An alternative approach is to assume that variation in behaviour is explained by variation in the model's parameters, either between or within subjects, or both. We will refer to this as the "Random Preference" approach. A third possible approach is to introduce what has come to be known as a "tremble term". This is a way of capturing misunderstandings and lapses of concentration. Sometimes, these three stochastic approaches are used in combination.

Perhaps the best illustration of all of these issues is in risky choice modelling (see Loomes et al., 2002; Harrison & Rutström, 2009; Conte et al., 2011; Von Gaudecker et al., 2011). The data set would typically consist of a sequence of binary choices between lotteries, performed by each of the subjects in a sample. Central to the modelling strategy is a von Neumann-Morgenstern utility function which captures the risk attitude (or "preferences") of an individual. Risk attitude clearly varies between individuals so the risk-attitude parameter takes the role of a subject-specific random effect. It is also clear that there is significant within-subject variation, and this is captured either by assuming that preferences vary over time for a given subject (the *random preference approach*), or that the subject makes a computational error each time he or she makes a choice (the *Fechner approach*). The

tremble assumption is a useful complement to each of these stochastic approaches since it allows for the occasional occurrence of extremely unlikely choices. An obvious theoretical framework on which to build these stochastic models is EU. Typically, however, EU is found to be over-restrictive, and models such as rank dependent (RD) theory (Quiggin, 1982) and cumulative prospect theory (Tversky & Kahneman, 1992) are found to fit the data better. These models include, in addition to risk aversion parameters, probability weighting parameters and (if the outcomes in the experiment include losses as well as gains) loss aversion parameters. The finite mixture approach is sometimes used to separate subjects into EU and RD "types" (Harrison & Rutström, 2009; Conte et al., 2011). All of these models can also be extended to allow certain parameters to depend on experience. This is useful for capturing the phenomena of, say, computational errors decreasing in magnitude with experience, or subjects moving closer to EU-maximization with experience (Loomes et al., 2002). All of these models may be estimated in a maximum likelihood framework, using maximisation routines that are computationally feasible.

## 1.7 Modelling Subject Heterogeneity

Perhaps the most important of all reasons for developing structural models in the context of experimental data is that they enable us to incorporate between-subject heterogeneity. There are two broad types of heterogeneity. *Discrete heterogeneity* is the situation in which the population of subjects is made up of a finite number of different "types" who respond in categorically different ways to stimuli. One example was mentioned at the end of the last sub-section: some individuals are EU-maximisers; others behave according to rank dependent theory (i.e. they weight probabilities). For another standard example, in the context of a public goods experiment, we might start by assuming that the population of subjects divides neatly into four types: "strategist", "altruist", "reciprocator", and "free-rider" (Bardsley & Moffatt, 2007).

*Continuous heterogeneity* is the situation in which subjects differ from each other in a dimension which is continuously measurable. A natural example is risk attitude. Every individual might be assumed to possess his or her own risk-aversion parameter and it is natural to assume that this varies continuously across the population.

The two types of heterogeneity call for different types of structural econometric models. Discrete heterogeneity calls for the use of finite mixture models, while continuous heterogeneity calls for the use of random-effects or random-parameter models. These two classes of model together play a central role in this book.

Of the "types" mentioned above that motivate the development of finite mixture models, probably the most important type in many situations is the "zero-type", that is, a subject who always contributes zero. In the context of a dictator game, such subjects may be labelled "selfish types", and in public goods games, they are labelled "free-riders". A very useful class of model that accounts for the existence of zero types is the "hurdle" framework. Hurdle models, or "double hurdle" models,

contain two equations, the first determining whether a subject is a "zero-type", and the second determining behaviour given that they are not a zero-type. Hurdle models play an important role in this book and, in particular, the hurdle framework is extended to panel data so that it is applicable to the situation typically arising in experiments where multiple decisions are observed for each subject. One reason why the hurdle framework is so useful is that it allows for the possibility that a treatment changes a subject's type, as well as just changing his or her behaviour. For example, there is a lot of current interest in whether the subject's endowment is earned or unearned. It is quite possible that if the endowment is a free gift, they are less likely to be a "zero-type" than if it is earned (which is normally the case in "real life"). If experimental conventions are found to change a subject's *type* from his or her real-life type (in addition to simply changing his or her behaviour), the implications are clearly important for the external validity debate. This is something that may be tested easily within the hurdle framework.

An important point about "subject types" is that we are not at any stage in a position to say with certainty that a particular subject is of a given type. For example, a subject observed contributing zero on every occasion in a public goods game is very likely to be a "free-rider", but we cannot say with certainty that he or she is a free-rider. The best we can do is to compute *posterior type probabilities* for each subject, following estimation of the model. This is done using Bayes' rule. The subject contributing zero every time would presumably have a very high posterior probability of being a free-rider. In a situation of continuous heterogeneity, we can use a similar technique. For example, in a risky-choice model in which we assume continuous variation in risk attitude between subjects, we can use Bayes' rule following estimation to obtain a posterior risk aversion estimate for each subject. This measure is useful in a number of ways.

In a finite mixture model, the likelihood contribution corresponding to a particular subject is a weighted average of probabilities or densities corresponding to the different types, with the type probabilities (or *mixing proportions*) as weights. In the presence of continuous heterogeneity, the estimation problem is somewhat more complicated because the likelihood contribution corresponding to a single subject becomes an integral over the variable(s) representing the heterogeneity. Hence, the procedure for evaluating the likelihood function must incorporate some numerical method for the evaluation of integrals. The method adopted throughout this text is the method of maximum simulated likelihood (MSL, see Train, 2003). This method is based on the principle that an integral can be computed by evaluating the function at each of a set of simulated values of the variable of integration, and then taking the mean of these function values. The simulated variables are not random numbers, but instead *Halton draws*, which result in greater efficiency in the evaluation of the integral.

## 1.8 Experimetrics of Other-regarding Preferences

As behavioural economics has taken hold, the economics profession has moved away from rigid assumptions of self-interested utility-maximisation and started to

introduce concepts such as "other-regarding preferences" and "inequity aversion". Nevertheless, it is often recognised that whatever the considerations that are guiding behaviour, they can somehow be incorporated into the utility function that the individual is assumed to be maximising. For example, the utility function may contain a component that represents self-interest, and resembles the traditional utility function, and a second component that represents how much importance the individual attaches to the welfare of those around him or her. Optimisation of such a function results in the "other-regarding" behaviour that many experiments have set out to investigate. An obvious way of doing this in the setting of, for example, a dictator game is to assume a utility function with two arguments, "own payoff" and "other's payoff", specify the utility function parametrically, and estimate the parameters using econometric techniques. This sort of estimation has been performed by Andreoni & Miller (2002) and others. Within the context of these models, the question of whether a treatment, such as whether the initial endowment is earned or unearned, has an impact on behaviour can be addressed by investigating the way in which it impacts on the parameters of the utility function, rather than simply by considering whether it impacts on the outcome variable. Jakiela (2013) carries out treatment tests of this type.

The chapter in this book that covers other-regarding preferences takes this type of utility function as the starting point and considers a variety of estimation methods. Specifically, an extension to the model is developed that incorporates zero observations on other's payoff (i.e. selfishness) in a theoretically consistent way, by treating them as corner solutions to the dictator's constrained optimisation problem. Another extension is a finite mixture model, similar to that of Cappelen et al. (2007), that assumes different subjects have different *fairness ideals*, and their behaviour is determined by this together with their degrees of selfishness. Finally, a completely different estimation strategy is introduced, that is suitable when the data consists of choices between different allocations (Engelmann & Strobel, 2004). The appropriate model is the conditional logit model. This model is found to be particularly useful for estimating the parameters of the well known Fehr & Schmidt (1999) utility function, which separately captures aversion to advantageous and disadvantageous inequality.

## 1.9 Experimetrics of Bounded Rationality

In interactive games, the fundamental prediction usually takes the form of a Nash equilibrium. The Nash equilibrium is based on the assumptions that agents hold correct beliefs about others' actions, and that agents best respond to these correct beliefs. As always, observed behaviour departs from the theoretical prediction and we need to consider ways of modelling such departures. The models used for this purpose are often thought of as models of bounded rationality. For a recent survey, see Crawford et al. (2013).

One such approach is the quantal response equilibrium (QRE) model (McKelvey & Palfrey, 1995), which assumes that each player's behaviour follows a distribution which is a "noisy" best response to other players' noisy behaviour.

Note that the best-response assumption is being relaxed, since decisions are "noisy". However, the correctness-of-beliefs assumption is satisfied, since players have correct beliefs about others' noisy behaviour.

A different approach is the level-k model (Nagel, 1995), which assumes that players have different (finite) levels of reasoning, with each player believing that all other players have a level of reasoning one below their own. Clearly, this assumption implies that, unlike in QRE, players have *incorrect* beliefs about others' behaviour. Closely related to the level-k model is the cognitive hierarchy model (Camerer et al., 2003) which assumes, perhaps more reasonably, that players believe that other players do not all have the same level of reasoning, but are instead distributed over levels of reasoning below their own.

The most obvious approach to operationalising the level-k and cognitive hierarchy models is to assume a zero-mean error around the deterministic best response at each level of reasoning, which amounts to a relaxation of the best-response assumption. This is the approach followed by Bosch-Domènech et al. (2010) and Runco (2013), and is also the approach followed in the relevant chapter of this book.

## 1.10 Experimetrics of Learning

If agents are found not to be operating exactly at the equilibrium, an obvious question arising is: is there a learning process by which they converge towards that equilibrium? This question may be addressed in settings in which each subject performs a sequence of tasks. The simplest way of addressing the question is to use the task number as an explanatory variable in a model for the decision variable. The effect of task number is then used to judge whether, and how quickly, behaviour is moving towards equilibrium as subjects gain experience.

Using the task number as an explanatory variable is all that is required in some situations. For example, in individual decision making, it is sometimes found that the parameters representing deviations from EU change with experience in a way that implies convergence towards EU (Loomes et al., 2002). It is also found that stochastic terms change with experience. In particular, the tremble probability is found to decay all the way towards zero in the course of an experiment, implying, reassuringly, that misunderstandings and complete losses of concentration are transitory phenomena. In public goods games in which a sequence of games are played, contributions are found to diminish with experience, and simply using the task number as an explanatory variable in the contribution equation is often found to be the best way of capturing this effect.

In those contexts, learning is (usually) only about the task, not about the behaviour of other players, nor even about the outcomes of previous tasks. A fairly standard feature of both of those settings is that, for reasons inherent to the experimental designs, subjects do not obtain feedback in terms of the outcomes of previous rounds. Hence, in those situations, the modelling of a learning process simply involves allowing particular parameters to depend in some way on the task number.

The situation in experimental games with feedback is very different. In each round, subjects 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 player's own past pay-offs and also the effects of past choices by other players. In econometric terms the modelling strategy moves from *static* to *dynamic* modelling since we now need to capture explicitly the relationship between current behaviour and past behaviour and outcomes.

A number of such models are considered in this book. *Directional learning* theory, first proposed by Selten & Stoecker (1986), is a simple form of dynamic learning model, in which subjects are assumed to adjust their behaviour in each period in response to the outcome of the previous period. *Reinforcement learning* (Erev & Roth, 1998) is based on the idea that a player's propensity to choose a strategy is a positive function of the *pay-offs* received as a result of choosing that strategy in previous periods. *Belief learning* (Cheung & Friedman, 1997) is based on the idea that players adjust their strategies in response to payoffs that *would* have been received under each choice.

The reinforcement learning model has been used primarily by psychologists, while the belief learning model has been used primarily by decision and game theorists. A model which nests these two models is the *experience weighted attraction* (EWA) model developed by Camerer & Ho (1999). This model is useful because it forms a framework for testing which of the two models better fits the data. However, the model contains a large number of parameters, and might be seen as over-parameterised.

## 1.11 What is in this Book?

The main objective of this book is to show the reader clearly and methodically how to carry out a wide range of tasks in Experimetrics. What this book does *not* contain is any detail in econometric theory, for example the derivation of the properties of estimators and tests. For these topics, the reader is referred to mainstream econometrics textbooks, such as Wooldridge (2012) or Greene (2008).

Another objective that has *not* been set for this book is to provide a comprehensive survey of the literature on each topic covered. In the main part of each chapter, the only citations will be to studies that are directly relevant to the techniques being covered. At the end of each chapter, there is a section containing a limited number of suggestions for further reading for the benefit of readers wishing to go further with particular topics.

Tasks are all demonstrated in STATA version 12 (StataCorp, 2011). All data sets used in the text are available online (www.palgrave.com/moffatt), and are listed in Appendix A. Some of the data sets used for demonstration are real and have already

been used in published work. Other data sets are simulated; please note that all simulated data sets contain the suffix **sim** in the name – it is important that if these data sets are used by readers, they are used only for practising techniques. Simulated data is clearly useful in situations in which suitable real data sets cannot be found. In fact, one of the chapters of the book is devoted to explaining how to simulate data sets with the required structure and features.

Many tasks can be carried out with existing STATA commands. Most of the STATA commands used in the book are listed in Appendix B. In certain situations, the required STATA commands do not exist, but user-written programs are available online which meet the requirement. These programs are found using the STATA `findit` command, from where they may be easily installed. For other tasks, programming in STATA is required. Basic STATA skills, such as how to create and run a do-file, are assumed, but the programs themselves will be explained in detail. In some situations MATA is used. MATA is a matrix programming language built into STATA. Some advanced tasks are only possible by including MATA commands within the STATA code.

In a small number of situations, Excel output is used, and some Excel files, also (www.palgrave.com/moffatt) are referred to. This is for particular types of problem that require flexibility in computation, for example in terms of outputs changing automatically in response to changes in inputs.

Exercises are included at the end of some chapters. The reason for the apparent lack of uniformity here is simply that some topics lend themselves more readily to exercises than others.

Chapter 2 considers the statistical aspects of experimental design that are most relevant to experimental economics. Most importantly, it provides a primer in "power analysis", the procedure used to select an adequate sample size for a treatment test. It also describes various types of design such as factorial, block, and within-subject designs, as well as explaining the difference between one-shot, partners, and strangers designs. Finally, it describes methods for administering multiple tasks per subject, namely the random lottery incentive scheme and the strategy method. Chapter 2 also introduces four very well-known experiments – the *ultimatum game*, the *dictator game*, the *trust game*, and the *public goods game* – which appear as applications many times throughout the book.

Chapter 3 covers treatment testing. This chapter contains many hands-on examples, some of which involve real data sets. It covers parametric and non-parametric tests, highlighting the strengths and weaknesses of the two approaches, and the circumstances under which each should be used. It also introduces the bootstrap, a method which makes a parametric test valid in a situation in which the underlying distributional assumptions do not hold. The chapter also covers tests comparing complete distributions and within-tests.

Chapter 4 considers treatment testing in the context of regression analysis. The applications are to data from auction experiments and contest experiments. A distinction is made between tests of the fundamental prediction of the theory, tests of comparative static predictions, and tests for the causes of out-of-equilibrium behaviour. The chapter includes methods for dealing with dependence, including clustering and the block bootstrap, and progresses to illustrations of panel data

estimation and multi-level modelling. Finally, an example of a meta-analysis applied to contest experiments is provided.

Chapter 5 presents a very different application of regression analysis, here to the analysis of decision times. This is an area that has rapidly become popular in recent years, partly because decision time is a useful measure of the effort expended by the subject (see for example Moffatt, 2005*b*). Panel data estimators and tests are once again illustrated in this context.

Chapters 6–7 are mainly concerned with modelling approaches that are appropriate when outcomes are discrete variables, for example binary data models, censored data models, interval regression, and ordinal models. In nearly all of these situations, built-in STATA commands are available for the task. In addition, in Chapter 6, the ML routine (Gould et al., 2010) in STATA is introduced and applied to simple maximum likelihood problems.

Chapter 8 introduces finite mixture models, with further use of the ML routine. The final example uses real data from a public goods experiment and presents the first application of the ML routine to a panel data problem.

Chapter 9 is devoted to simulation of experimental data. This includes sections on how to simulate panel data, dynamic panel data, and binary panel data. As mentioned, simulated data is useful in situations in which real data is unavailable. Simulated data is also useful for the testing of programs, and for investigating the properties of estimators and tests. The latter is done using the Monte-Carlo technique. The use of Monte-Carlo is demonstrated in the chapter, including an application to the problem of evaluating the performance of a test statistic developed in Chapter 7.

Chapter 10 introduces the method of maximum simulated likelihood (MSL) which is adopted as the standard modelling framework for dealing with continuous heterogeneity. This chapter includes examples of MSL being applied to simulated data sets. Chapter 11 contains the first application of MSL to the estimation of the panel hurdle model, a model which allows for the presence of a "zero-type" in panel data settings. The panel hurdle model is applied to real data sets from previously published work. One strength of the hurdle framework that is emphasised is that it allows a treatment to change a subject's type in addition to altering his or her behaviour. Chapter 12 covers theoretical issues relating to choice under risk, preparing the ground for Chapter 13 which covers the econometric modelling of risky choice models. This is the second application of the MSL method.

Chapter 14 is concerned with optimal design of binary choice experiments. This chapter draws on the well-developed theory of optimal design from the statistics literature, and applies it to a particular type of experiment in economics.

Chapter 15 is concerned with the estimation of social preference models. The focus here is the estimation of the parameters of a utility function whose arguments are "own payoff" and "other's payoff". Real data from a dictator game are used for illustration. The parameters of the utility function are estimated using various different approaches, including: a model with binding non-negativity constraints to explain zero observations; a mixture model allowing individuals to differ in the nature of their fairness ideal; and a discrete choice model that uses data on subjects' choices between allocations.

Chapters 16–18 are concerned with the econometric modelling of data from experimental games. Chapter 16 explains the quantal response equilibrium (QRE) model, and illustrates its estimation using real data from the pursue-evade game. The QRE model is also applied to the contest data from Chapter 4. Chapter 17 develops the level-k and cognitive hierarchy models, explaining how they are estimated using simulated data from a guessing game. Chapter 18 covers a number of models of learning: directional learning (DL); reinforcement learning (RL); and belief learning (BL). It ends with an explanation of the experience weighted attraction (EWA) model, which is a heavily parameterised model nesting both RL and BL.

An important feature of the book is the close linkages between chapters. Because this is a subject area in which themes tend to emerge in different contexts, there is a good deal of cross-referencing between chapters. For example, the data set used in Chapter 13, on the econometric modelling of choice under risk, is simulated using the techniques described in Chapter 9, and estimation is conducted in the framework of MSL which is explained in detail in Chapter 10. Also, one of the by-products from the estimation of the risky-choice model in Chapter 13 is a measure of "closeness-to-indifference" which plays a very important role in the model of decision times estimated in Chapter 5.

Finally, there are three appendixes. Appendix A provides a list of the data sets and other files that are referred to in the book, and are available online at www.palgrave.com/moffatt. Appendix B provides a list of most of the STATA commands used in the book, with a brief explanation of each. Appendix C contains a table defining the 50 choice problems used in the (simulated) risky choice experiment analysed in Chapters 5 and 13.

# Chapter 2
# Statistical Aspects of Experimental Design in Experimental Economics

## 2.1 Introduction

This chapter is concerned with aspects of experimental design in experimental economics. In experimental economics, unlike in most other areas of economics, the data generating process is governed by the analyst. Because of this, assumptions required for the identification of a treatment effect are much less stringent than in other areas. In an experimental setting, the only major assumption required for identification is appropropriate randomisation (with appropriate sample sizes). As pointed out by List et al. (2011), randomisation plays the role that an instrumental variable would play in a situation of naturally occurring data.

Issues of experimental design are relatively simple if the framework is one of treatment testing. One central design issue is the required sample size. A useful framework for choosing a sample size is power analysis. The principal objective is to find the sample size that is required to attain a pre-set level of power for a given treatment test. This chapter includes a primer on power analysis.

Other design issues include matching methods: partners versus stranger matching. This is closely related to the issue of clustering. We will also consider designs such as the Random Lottery Incentive (RLI) system and the Strategy Method (SM).

## 2.2 The Average Treatment Effect

This section provides a formal framework for the analysis of treatment effects.

It is usually recognised that every individual has his or her own treatment effect. Also, it is usually assumed that individual treatment effects vary randomly around an average. We are principally interested in estimating the *average treatment effect* (ATE).

Consider the effect of a particular treatment on a particular outcome variable $Y$. Let $T$ be a binary variable representing treatment status: $T = 1$ for treatment; $T = 0$ for control. Let $Y_i(T)$ be the outcome for subject $i$ given treatment status $T$.

We assume the following simple model for the outcome variable:

$$Y_i(T) = \alpha + \beta' X_i + \bar{\tau} T + \tau_i T + \epsilon_i \tag{2.1}$$

$X_i$ is a vector of observed individual characteristics (such as gender) which are thought to affect the outcome; $\bar{\tau}$ is the average treatment effect (ATE); $\tau_i$ is the subject-specific treatment effect, where $E(\tau_i) = 0$; $\epsilon_i$ is an independent and identically distributed (i.i.d.) random error term. The ATE may then be defined as:

$$\bar{\tau} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \tag{2.2}$$

A problem that is central to treatment testing is that we cannot, in general, observe the two quantities appearing on the right hand side of (2.2). For a given $i$, we can only observe the quantities $E[Y_i(1)|T = 1]$ and $E[Y_i(0)|T = 0]$; that is, it is only possible to observe average behaviour under the treatment for those whom we choose to treat, while it is only possible to observe average behaviour without treatment for those whom we choose not to treat. If the propensity to receive treatment is correlated with observed or unobserved characteristics of the subject, the estimate $(\hat{\tau})$ of the ATE will be biased, since:

$$\hat{\tau} = E[Y_i(1)|T = 1] - E[Y_i(0)|T = 0] \neq E[Y_i(1)] - E[Y_i(0)] = \bar{\tau} \tag{2.3}$$

Randomisation is used to ensure that assignment to treatment is independent of other sources of variation, so that $E[Y_i(1)|T = 1] = E[Y_i(1)]$ and $E[Y_i(0)|T = 0] = E[Y_i(0)]$, giving equality in (2.3), implying that the estimated treatment effect is unbiased for the ATE.

A key assumption implicit in this framework is that the distribution of subject-specific treatment effect ($\tau_i$ in (2.1)) is "well-behaved", that is, has a bell-shaped and symmetric distribution arround the ATE. In certain situations, it is reasonable to expect this assumption to fail. For example, it may be that within the population of subjects, half respond to the treatment with a treatment effect of $+1.0$, while the remaining half are irresponsive to the treatment, and therefore have a treatment effect of zero. The ATE in this situation would obviously be $+0.5$ but this is a misleading measure of the effect of the treatment, since it is not close to the actual treatment effect of any individual subject. The best way to deal with such discreteness in the distribution of treatment effects is to apply the mixture modelling framework (McLachlan & Peel, 2000), in which it is assumed that there is more than one "subject type", with types differing from each other in the way in which they respond to the treatment, and with the proportion of each type in the population, the "mixing proportions", being estimated as additional parameters. The mixture modelling approach is one that is used many times in later chapters of this book.

## 2.3 Randomisation Techniques

As implied in the last section, the importance of randomisation is that it gives rise to a situation in which identification is not a problem. Here, we provide a discussion

of popular randomization techniques. For more detail the reader should consult List et al. (2011).

### 2.3.1 Completely randomised designs

The simplest of experimental designs is the completely randomised design. A random sample is drawn from the entire subject pool, and treatments are probabilistically assigned to subjects, independently of any observed or unobserved characteristics. The advantage of this procedure is that, by definition, it minimises the risk that treatment is correlated with subject characteristics. The disadvantages are that the sample sizes in each treatment are random, and the variance of the outcome may be large. Both of these factors tend to reduce the ability of the analyst to draw statistical inferences from the experimental data.

### 2.3.2 Factorial designs

An obvious solution to the problem of random sample sizes arising with the completely randomised design is to assign pre-determined numbers of subjects to each treatment, or to each combination of treatments. However, it is important that subjects are not assigned to treatments in the order in which they arrive, since time of arrival is likely to be correlated with subject characteristics. On arrival subjects should be given a random number determining treatment assignment, and recruitment should cease when all of the pre-determined targets are met.

Consider the following example of a giving experiment (e.g. a dictator game). Let us assume that there are two treatments: high stakes versus low stakes (with low stakes treated as "control"); and communication (between the two players) versus no communication (with no communication treated as "control"). Let us assume that we design the experiment with the numbers appearing in the following table; that is, with each *combination* of treatments being applied to 30 subjects, and a total of 120 subjects. Each cell in the table represents a "trial".

| | Low stakes | High stakes |
|---|---|---|
| No communication | 30 | 30 |
| Communication | 30 | 30 |

This is known as a "full-factorial design", because all possible treatment combinations are being covered. It might also be referred to as a "2 × 2 design". This sort of design is useful if, in addition to the two average treatment effects, we are interested in the "interaction" between the two treatments. For example, we might hypothesise that communication is less important when the stakes are higher, perhaps because financial incentives "crowd out" intrinsic motivation. In order to test such an interaction effect between the two treatments, the full-factorial design would be necessary.

However, if we are not interested in such an interaction effect, and only interested in the "main effects", that is, the effects of the treatments themselves, then the following design would suffice:

| | Low stakes | High stakes |
|---|---|---|
| No communication | 30 | 30 |
| Communication | 30 | 0 |

Here, only three trials are used, and the total number of subjects is only 90. This is known as a "fractional factorial design". If there were a larger number of treatments, the difference between full-factorial and fractional-factorial would become more substantial. To be precise, if there are $m$ different treatments, the full-factorial design would require $2^m$ trials, while a fractional-factorial design which identifies all $m$ main effects would only require $m + 1$ trials. Clearly large savings in the sample size are possible in situations in which only main effects are of interest.

### 2.3.3 Block designs

If the subject pool is heterogeneous in certain observable dimensions, it may be advantageous to apply a block design. Experimental units are divided into blocks in accordance with the observable characteristics (contained in the vector $X_i$ in (2.1)). Then randomisation is performed within, but not between, blocks. The variable on which blocking is applied is known as the blocking factor. Typically, a blocking factor is a source of variability that is not of primary interest to the experimenter.

One obvious choice of blocking factor is gender. Gender may be an important source of variability in the outcome, and by blocking on it, this source of variability is controlled for, leading to greater accuracy in the estimation of the treatment effect(s) of central interest.

### 2.3.4 Within-subject designs

A within-subject design (or repeated measures design) can be thought of as a special case of a block design in which the experimenter blocks on a single subject, and the subject experiences more than one treatment. The considerable advantage of within-subject designs is that the impact of the subject-specific effect ($\alpha_i = \alpha + \epsilon_i$ from (2.1)) is essentially eliminated, and this has the potential to improve greatly the precision of the treatment effect estimate.

Formally, if $\hat{\tau}_{bs}$ is the between-sample estimate and $\hat{\tau}_{ws}$ is the within-sample estimate obtained using the same total number of observations, $N$, then (List et al., 2011):

$$V(\hat{\tau}_{ws}) = V(\hat{\tau}_{bs}) - \frac{2}{N} V(\alpha_i) \tag{2.4}$$

The interpretation of (2.4) is that if all subjects are identical, so that $V(\alpha_i) = 0$, there is no benefit from using a within-sample design, but if subjects differ greatly, so that $V(\alpha_i)$ is large, the benefits of the within-subject design are considerable.

A disadvantage of the within-subject design is the possibility of "order effects"; that is, the behaviour of subjects depending on the order in which the treatments are experienced.

### 2.3.5 Crossover designs

The problem of order effects may be addressed by varying the order of treatments between subjects. For example, if there are two treatments A and B, half of the subjects may be assigned to the sequence AB, and the other half to the sequence BA. Differences between these two groups would confirm the existence of an order effect, which would then need to be controlled for in treatment tests.

### 2.3.6 ABA designs

"ABA" refers to a design which starts 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 so the behaviour can be observed a second time (A). This makes it possible to measure behaviour before treatment, during treatment, and once treatment is removed.

## 2.4 How Many Subjects? A Primer in Power Analysis

In deciding how many subjects to recruit, it is often useful to make use of power analysis (Cohen, 2013). Power analysis is the name given to the formal process for determining the sample size for a research study. The essence of the process is that it determines the sample size necessary for the test to achieve a specified *power*. The power of a test is the probability of detecting a "true" effect when a true effect actually exists.

The power analysis used in this section is based on the assumption that the outcome is a continuous. In Chapter 14, we tackle the perhaps more demanding problem of optimal design in a situation in which the outcome is binary.

### 2.4.1 The case of one sample

Suppose that we are interested in the continously distributed outcome measure $Y$ whose population mean is $\mu$. Suppose further that we are interested in testing the

null hypothesis $\mu = \mu_0$ against the alternative hypothesis $\mu = \mu_1$, where $\mu_1 > \mu_0$.[1] We plan to collect a sample of size $n$ for this purpose, and we need to decide what $n$ should be. Before we do this, we need to set two quantities. The first is the test size, $\alpha$, 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 $\beta$. 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 $\pi$.

It has become standard to set $\alpha$ to 0.05, unless there are compelling reasons to do otherwise. Although there are no formal standards for power, many researchers assess the power of their tests using $\pi = 0.80$ as a standard for adequacy. The corresponding value of $\beta$ is 0.2. These conventions imply a four-to-one trade off between the probability of type II error and the probability of type I error.

Having decided on these values of $\alpha$ and $\beta$, we proceed to apply power analysis. The test that will be performed is the one-sample t-test, which is based on the following test statistic:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \tag{2.5}$$

where $\bar{y}$ and $s$ are respectively the mean and standard deviation of the sample which is of size $n$. Under the null hypothesis, $t$ defined in (2.5) has a $t(n-1)$ distribution. Hence, the rejection rule, given our chosen value of $\alpha$, is $t > t_{n-1,\alpha}$.

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 $t > z_\alpha$. This simplifies the analysis considerably.

The power of the test is given by:

$$\begin{aligned} P(t > z_\alpha | \mu = \mu_1) &= P\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}} > z_\alpha \mid \mu = \mu_1\right) \\ &= P\left(\bar{y} > \mu_0 + \frac{z_{0.05}}{s/\sqrt{n}} \mid \mu = \mu_1\right) \\ &= P\left(\frac{\bar{y} - \mu_1}{s/\sqrt{n}} > \frac{\mu_0 + z_\alpha s/\sqrt{n} - \mu_1}{s/\sqrt{n}} \mid \mu = \mu_1\right) \\ &= \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha s/\sqrt{n}}{s/\sqrt{n}}\right) \end{aligned}$$

[1] Alternative hypotheses nearly always involve inequalities, for example, $\mu > \mu_0$ or $\mu \neq \mu_0$. 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 of finding the desired sample size to be properly defined. The value under the alternative is assumed to derive either from prior beliefs, from a previous study, or from a pilot study.

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

$$\frac{\mu_1 - \mu_0 - z_\alpha s/\sqrt{n}}{s/\sqrt{n}} = z_\beta \tag{2.6}$$

Rearranging (2.6) we obtain:

$$n = \frac{s^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}$$

Recalling that our chosen values of $\alpha$ and $\beta$ are 0.05 and 0.20 respectively, we have $z_\alpha = 1.645$ and $z_\beta = 0.84$. Hence we may write the formula for the required sample size as:

$$n = \frac{6.17s^2}{(\mu_1 - \mu_0)^2} \tag{2.7}$$

For an example of the use of the formula (2.7), suppose that we are testing the null $\mu = 10$ against the alternative $\mu = 12$, and we happen to know that the standard deviation of the data is 5. Then we apply formula (2.7) to obtain:

$$n = \frac{6.17 \times 5^2}{(12 - 10)^2} = 38.6$$

Clearly $n$ needs to be an integer, and in order to ensure that the power requirement is met (i.e. that the power is *at least* 0.8), we should round up rather than down. The required sample size in this example is therefore 39.

The STATA command `sampsi` can be used to perform this calculation directly. This is an example of one of STATA's "immediate" commands. An immediate command (which always ends with the letter i) is a command that obtains results not from the data stored in memory but from numbers typed as arguments. To conduct the analysis for the above example using the `sampsi` command, the required syntax is:

```
. sampsi 10 12 , sd(5) onesam oneside p(0.8)
```

The main arguments are the values under the null and alternative (10 and 12). The options are as follows: "sd(5)" indicates that the known standard deviation is 5; "onesam" indicates that a one-sample test is required; "oneside" indicates that a one-sided test is required; "p(0.8)" indicates that the required power is 0.8. The output from this command is as follows. Note that the required sample size is 39, in agreement with the calculation performed above.

```
Estimated sample size for one-sample comparison of mean
  to hypothesized value
Test Ho: m =  10, where m is the mean in the population
Assumptions:
          alpha =   0.0500  (one-sided)
          power =   0.8000
```

```
alternative m =       12
           sd =        5
Estimated required sample size:
            n =       39
```

## 2.4.2 *Choosing the sample size in a treatment test*

We now consider the slightly more complicated situation that is more usual 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. Again power analysis can be used to determine the sample size that is required to meet this objective.

Let $\mu_1$ and $\mu_2$ be the population means of the control group and the treatment group respectively. The null hypothesis of interest is $\mu_2 - \mu_1 = 0$ (i.e. the treatment has no effect), and the alternative is $\mu_2 - \mu_1 = d$ (i.e. the treatment has an effect of magnitude $d$). $d$ 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 $d$ is assumed to be derived either from prior beliefs, from a previous study, or from a pilot study.

The testing procedure that is required to test the null hypothesis $\mu_2 - \mu_1 = 0$ is the independent samples t-test. If the two sample sizes are $n_1$ and $n_2$, the sample means are $\bar{y}_1$ and $\bar{y}_2$, and the sample standard deviations are $s_1$ and $s_2$, the independent samples t-test statistic is given by:

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

where $s_p$ is the "pooled" sample standard deviation and is given by:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

The distribution of $t$ under the null hypothesis is $t_{n_1+n_2-2}$. Again, matters are simplified using the normal approximation. We will therefore use the critical value $z_\alpha$.

In this two-sample situation, we clearly need to find two required sample sizes, $n_1$ and $n_2$ say, one for each sample. However, we start by constraining the two sample sizes to be equal, that is, $n_1 = n_2 = n$. The test statistic becomes:

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

The power of the test is given by:

$$P(t > z_{0.05} | \mu_2 - \mu_1 = d) = P\left( \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{2}{n}}} > z_\alpha \middle| \mu_2 - \mu_1 = d \right)$$

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

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

$$= \Phi\left( \frac{d - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}} \right)$$

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

$$\frac{d - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}} = z_\beta \tag{2.8}$$

Rearranging (2.8) we obtain:

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

Once again applying our chosen values of $\alpha$ and $\beta$, we have $z_\alpha = 1.645$ and $z_\beta = 0.84$, and we may write the formula for the required sample size as:

$$n = \frac{12.35s_p^2}{d^2} \tag{2.9}$$

For an example of the use of formula (2.9), suppose that we are testing the effect size $d = 2$, and we know that the standard deviations of populations 1 and 2 are 4.0 and 5.84 respectively. Given that the two sample sizes are constrained to be equal, the pooled standard deviation is 5.0. Then we apply formula (2.9) to obtain:

$$n = \frac{12.35 \times 25}{4} = 77.2$$

Rounding up, we arrive at the required sample size (in each treatment) of 78.

The STATA syntax for the test just performed is:

```
. sampsi 10 12 , sd1(4.0) sd2(5.84) oneside p(0.8)
```

The main arguments are the values of $\mu_1$ and $\mu_2$. We could use any values here, provided their difference is 2 (the effect size). The options are the two standard deviations, and the request for a one-sided test. The output from this command is shown below. Note that the required sample size is in agreement with the calculation performed above.

```
Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.8000
            m1 =       10
            m2 =       12
           sd1 =        4
           sd2 =     5.84
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =       78
            n2 =       78
```

### 2.4.3 *Treatments with unequal costs*

In the above analysis, there was a constraint that the sample sizes were equal across the two treatments. One reason for relaxing this constraint would be a difference in the sampling costs between treatment and control. Suppose the focus of the experiment is on the effect of incentives, and there is a low incentive treatment and a high incentive treatment. It is logical that costs are expected to be higher in the latter treatment.

In this situation, in order to attain a desired level of power while respecting a cost constraint, the sample sizes should be set so that the ratio of the sample sizes is proportional to the square root of the cost ratio. Specifically, if the cost-per-subject in the two treatments are $c_1$ and $c_2$, it should be the case that:

$$\frac{n_2}{n_1} \propto \sqrt{\frac{c_1}{c_2}} \tag{2.10}$$

For example, let us suppose the cost per subject in the high-incentive treatment is four times as high as that in the low-incentive treatment. Applying (2.10), we obtain that there should be twice as many subjects in the low-incentive treatment, as in the high-incentive treatment.

This requires one more option (r(.)) in the sampsi command. With this option, the ratio $n_2/n_1$ is specified. If the low-incentive treatment is treatment 1, we here require $r = 0.5$. The command and output are shown below. Note that, as required, the desired low-incentive sample is (almost exactly) twice as large (131) as the desired high-incentive sample (66).

```
. sampsi 10 12 , sd1(4) sd2(5.84) oneside p(0.8) r(0.5)

Estimated sample size for two-sample comparison of means Test Ho: m1
= m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.8000
            m1 =       10
            m2 =       12
           sd1 =        4
           sd2 =     5.84
         n2/n1 =     0.50

Estimated required sample sizes:

            n1 =      131
            n2 =       66
```

Of course, if the experimenter really is working subject to a cost constraint, and the desired sample sizes turn out to be too large for the budget, it becomes necessary to relax the power requirement. For example, reducing the power from 0.80 to 0.60, we obtain considerably lower required sample sizes, which may be within the budget constraint:

```
. sampsi 10 12 , sd1(4) sd2(5.84) oneside p(0.6) r(0.5)

Estimated required sample sizes:

            n1 =       76
            n2 =       38
```

### 2.4.4 Sample size in cluster designs

Consider the following situation. You teach a module on which there are 300 students. The students are divided into 30 "seminar groups", each of size 10, which meet weekly. You wish to conduct an experiment and you wish to select the sample of subjects from the students taking the module. Let us suppose that you have used power analysis along the lines of the previous sub-sections, and you have determined that 60 subjects are required for your experiment. It would be administratively convenient randomly to select 6 of the 30 seminar groups, and to perform the experiment on the 60 students in the selected groups. This would be a "cluster design", in which the seminar groups are the "clusters".

Unfortunately, set against the convenience of the cluster design, there is a complication. If, as we might expect, outcomes are correlated within groups, a larger sample is required to obtain the desired level of power. To formalise this, we assume the following model, in which $u_j$ is a group-specific error term for group $j$:

$$Y_{ij}(T) = \alpha + \bar{\tau}T + u_j + \epsilon_{ij}$$

To attain the desired power, the sample sizes determined on the assumption of independent observations need to be inflated by the following factor (see List et al. (2011)):

$$1 + (c - 1)\rho$$

where $c$ is the size of each cluster (10 in the example), and $\rho$ is the "coefficient of intracluster correlation", defined as:

$$\rho = \frac{var(u_j)}{var(u_j) + var(\epsilon_{ij})} \tag{2.11}$$

To understand equation (2.11), first imagine that there are no differences between groups, so that $var(u_j) = 0$ and therefore $\rho = 0$. The inflation factor (2.4.4) is then 1, meaning that the required sample size is the same as before. Next, imagine that there are differences between groups, and furthermore that all members of a group behave identically to each other, so that $var(\epsilon_{ij}) = 0$, and therefore $\rho = 1$. Then, again using (2.4.4), the sample size would need to rise by a factor $c$ (group size). This is because, in this extreme situation, sampling more than one subject from within a group is worthless, and the sample-size requirement simply becomes a requirement on the number of clusters.

In practice, the value of $\rho$ that we expect is a small positive number, representing modest intergroup differences. For example, if $\rho = 0.05$ in the example, the sample size would need to rise by a factor 1.45, that is, from 60 to 87. This would mean that we would need to increase the number of sampled seminar groups from 6 to 9.

## 2.5 Four Very Popular Experiments

Four of the most popular experimental settings are the ultimatum game, the dictator game, the trust game, and the public goods game. These, or combinations of them, appear many times as applications throughout the text. For this reason, we describe them in this section.

### *2.5.1 The ultimatum game*

The *ultimatum game*, introduced by Güth et al. (1982) is described as follows. Two players ("proposer" and "responder") bargain over the division of some fixed money amount ($100 say). The proposer moves first by offering some split of the pie (e.g. "I get $65; you get $35"). The responder then has the choice between two actions: they either "accept", in which case the proposed division is implemented; or they "reject", in which case both players receive zero. There is a unique subgame-perfect equilibrium in which the proposer offers a split that gives the responder the smallest allowable payoff (perhaps $1), and the responder accepts. There is extensive experimental evidence that behaviour deviates from the subgame-perfect equilibrium in predictable ways: proposals are often very generous, and sometimes specify an even split; responders frequently reject "low" offers. Camerer (2003) reports the overall

findings that proposers offer an average of 40% of the money, and "low" offers of around 20% are rejected about half the time.

### 2.5.2 The dictator game

The *dictator game* is a simplified version of the ultimatum game. In the dictator game, the proposer again determines an allocation of some fixed money amount. However, the responder simply receives the remainder left by the proposer. The responder's role is entirely passive; he has no strategic input towards the outcome of the game. It is sometimes said that the responder has no "power of veto" in the dictator game. 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, proposers would allocate the entire money amount to themselves and would give nothing to the responder. There is much experimental evidence to reject this model: a significant proportion of dictators are observed giving positive amounts to the responder. On average, dictators give about 20% of the endowment to the responder (Camerer, 2003).

### 2.5.3 The trust game

In the *trust game* (Berg et al., 1995), the two players are "sender" and "recipient". The sender has an endowment, and he or she has the opportunity to send some or all of it to the recipient. The experimenter then (typically) triples the amount sent. After the recipient receives the transfer (i.e. three times the amount sent), he or she may return money back to the sender. The amount sent by the sender is a natural measure of "trust", while the amount returned by the recipient is (after controlling for the amount sent) a natural measure of "trustworthiness".

The trust game is related to the dictator game in the sense that the recipient is a dicator whose endowment has been given to him or her by the sender.

In the trust game thus decribed, the unique subgame-perfect equilibrium is such that the sender sends zero to the recipient. This is because the sender has no reason to send any positive amount given that they expect the recipient to send zero in return. Once again, experimental evidence departs from this prediction in more than one way. According to the meta-analysis of Johnson & Mislin (2011), senders send on average around 50% of their endowment, signalling a modest degree of trust, while recipients return around 37% of the (tripled) amount received, which is (just) sufficient for the trust to "pay".

### 2.5.4 The public goods game

A public goods experiment is typically conducted using what is known as the voluntary contributions mechanism (VCM), outlined as follows. Subjects are divided

into groups of $n$ members. Each group member has an endowment of $E$ tokens, that they must divide between a private account and a public account. Each token that a subject allocates to her own private account earns one point for her (and nothing for anyone else); in contrast, each token that any group member allocates to the public account is multiplied by $m$ and then divided equally among all $n$ group members. Hence, for each token the group member allocates to the public account, *every* group member earns $m/n$ units. This ratio is called the marginal per capita return (MPCR). It is usually the case that $n > m > 1$, so that the MPCR, though positive, is strictly less than unity.

The game has a unique Nash equilibrium consisting of zero contributions by every subject. This is obvious when considering that while the subject earns one unit when allocating a token to her private account, she earns an amount $m/n < 1$ when allocating the same token to the public account. Hence, regardless of the allocations of other group members, each subject maximises her own pay-off by allocating all of her endowment to her private account.

It is also obvious that this Nash equilibrium is socially inefficient, as it yields $E$ units per group member when, had all members contributed their full endowment to the public account, each would have received $mE > E$ units. Note that if the group consisted of only two players, the situation would be very similar to the well-known prisoner's dilemma.

Experiments using the VCM procedure described above have been surveyed by Ledyard (1995). One overall finding is that the average subject contributes around 40% of his or her endowment to the public account. When analysing data from such games, our primary interest will be in identifying the motivations behind such contributions. One motivation to which we shall pay particularly close attention is the reciprocity motive: subjects' contributions depending positively on previous contributions by others. Another phenomenon of interest will be the widely observed decay in the level of contributions as the game is repeated.

## 2.6 Other Aspects of Design

### *2.6.1 The random lottery incentive (RLI) mechanism*

The random lottery incentive (RLI) mechanism is an elaborate form of within-subject design. Each subject completes a number of distinct decision tasks. The tasks might be valuations of lotteries, choices between pairs of lotteries, or choices of strategy in a game. Each task typically has a well-defined reward structure where the payoff is a function of the choice made, and also of the moves of nature (in games of chance), or of the moves of other players (in games of strategy). At the beginning of the experiment, the subject is made aware that when the sequence of tasks is completed, *one* of the tasks will be selected at random, and the payoff to the subject will be the outcome of the selected task.

Since only one task will be for real, the RLI is likely to encourage subjects to think about each task as if it were for real, and as if it were the only task faced. If subjects do think in this way, then the mechanism will have the desirable consequence of eliminating wealth effects that could arise if payoffs from tasks influence subsequent decisions.

The RLI may also be administered as a form of "crossover" design if the order of tasks is varied between subjects. This is very useful for identifying the role of experience in decision making.

### 2.6.2 The strategy method

The strategy method, introduced by Selten (1967), is a procedure for eliciting subjects' responses in games. The ultimatum game provides a useful context in which to describe this method. In standard implementations of the ultimatum game, the data generated consists of *observed decisions*; that is, the proposer's offer and the responder's response are recorded. This is known as the "direct decision approach". Notice that a limitation of this approach is that it only reveals the responder's decision and no more: we observe their response to the proposer's actual offer, but we do not observe how they would have responded to different offers that could have been made.

The basic principle of the strategy method is to ask each player to reveal their entire *strategy*. This requires the responder to give a conditional response to each offer that could be made, before knowing what offer has actually been made. Once the strategy has been elicited, the game can be played out by implementing the strategy.

The obvious advantage of the strategy method is that the information obtained is richer since it contains information about how agents would behave at information sets that might not arise in the actual course of play. This is particularly relevant in the case of the ultimatum game because the strategy method allows us to observe how agents respond to very low offers, despite low offers being very rare.

### 2.6.3 "One-shot", "partners", and "strangers" designs

Theories in their purest form tend to assume that the game is played only once, and also that players are fully rational with correct beliefs, and do not require experience of the game. If we are really only interested in subjects' behaviour in a single play of the game, we simply ask them to play one time against opponents randomly selected from the pool of subjects present in the current session. This design may be labelled the "one-shot" design.

However, in game theory experiments, it is more usual for subjects to play repeatedly over a sequence of rounds, for the obvious reason that this generates more data, and also because it provides an opportunity for subjects to gain experience. The

reason why we are keen for subjects to gain experience is that we are often more interested in the behaviour of experienced subjects than that of inexperienced subjects. We are also interested in the process by which the subjects gain experience, that is, the *learning* process.

The question which then arises is how exactly the repetition should be administered. There are two popular types of protocol. In *partners* designs (sometimes called *fixed rematching designs*), the same group of subjects play together in every round. In this situation, there are two possible reasons why behaviour might change between rounds: learning (i.e. subjects require experience in order to gain an understanding of the incentives of the game); and strategic considerations (i.e. a subject's behaviour changes as she updates her beliefs about other players in the group). The other type of protocol is the *strangers* design (sometimes called the *random rematching design*). Here, the groups who play together are randomly reselected (from the pool of subjects present in the current session) each round. As suggested by Andreoni (1988), the purpose of the strangers design is to separate the effect of learning from that of strategic considerations. If behaviour changes between rounds under a strangers design, it is reasonable to attribute the change to learning alone, since strategic considerations are absent.

There are different forms of strangers design, according to whether restrictions are imposed on the random process that reselects groups. The term *perfect strangers* is often used for a design that restricts the random process in such a way that the probability of any two players playing each other twice is zero. However, this leaves scope for the possibility that player $i$ anticipates that a player ($j$) that she is playing with in round $t$ will meet another player ($k$) in round $t+1$, and that she ($i$) might meet $k$ in round $t+2$. Hence it is conceivable that subjects make choices on the basis that these choices might indirectly influence the behaviour of subjects they will meet in future rounds. Arguably, the term *perfect strangers* should be reserved for designs in which, in each round, each subject $i$ plays subjects who have not previously met $i$, or any of those whom $i$ has met, or any of those whom they have met, and so on. Such a design would completely eliminate the possibility of player $i$'s choices in round $t$ influencing the choices of players with whom $i$ plays in future rounds. However, such a design is considerably more complex and costly to implement than unconstrained random rematching.

## 2.7 Summary and Further Reading

A key concept covered in this chapter is randomisation. This, and other issues central to experimental design in economics, is covered in detail by List et al. (2011) and Green & Tusicisny (2012).

The point was made early in the chapter that randomisation is important because it gives rise to a situation in which identification is not a problem. However, it should be recognised that a problem that remains is generalisability (also known as external validity). According to Al-Ubaydli & List (2013), there is a tradeoff between identification and generalisability. Some popular randomisation techniques

have been discussed. For more detail on these the reader should consult List et al. (2011).

Power analysis has also been explained in some detail in this chapter. For further detail, the reader is referred to Cohen (2013).

For more information on the "popular experiments" described in Section 2.5, and also on the various aspects of experimental design covered in Section 2.6, the reader is referred to the relevant sections of Bardsley et al. (2009).

## Exercises

1. Burnham (2003) considers a "photograph" treatment in a dictator game. Given the effect sizes reported, find the optimal sample sizes.
2. What assumptions are required for the random lottery incentive system to be incentive compatible?

# Chapter 3
# Treatment Testing

## 3.1 Introduction

The last chapter was concerned with issues of experimental design in the context of treatment testing. The most important design feature was the choice of sample size in treatment testing. This chapter is concerned with methods for implementing treatment tests, having implemented the design and collected the data.

There are two broad approaches to treatment testing. The first is the "between-subject" approach, in which the sample is divided into two groups, the treatment group, to whom the treatment is applied, and the control group, to whom no treatment is applied. An outcome measure is recorded for each subject. The other approach is the "within-subject" approach. This is to obtain the outcome measure from each member of the sample, both without and with the treatment. Whichever of these two approaches is followed, the central question that is always being addressed is whether the treatment influences the outcome, and if so, in which direction.

Having decided between the "within-subject" and the "between-subject" approaches, there are then many different ways in which the test can be performed. Another broad division is between parametric and non-parametric tests. Both types of test have advantages and disadvantages.

The key factor in the choice between non-parametric and parametric tests is the scale of measurement of the data. There is a large literature on this point (see, for one example, Harwell & Gatti, 2001). There are essentially three scales of measurement: nominal, ordinal, and cardinal. Parametric tests, to an extent, rely on distributional assumptions which can only hold if the variables in question are measured on a cardinal scale.

Even if measurement is on a cardinal scale, some experimental economists seem uncomfortable with the use of parametric tests, because they worry that the distributional assumptions may not be met. There are two important responses to this concern. Firstly, provided that the number of observations in each treatment is sufficiently large, it is possible to appeal to the central limit theorem (CLT) which, under certain conditions, implies that the (standardised) mean of a sample follows a normal distribution even when the sample is drawn from a distribution that is not normal (see e.g. Berenson et al., 1988). Secondly, even in a situation in which

the CLT cannot be relied upon (e.g. low sample sizes), a method is available for ensuring that inferences made on the basis of parametric tests are valid regardless of the distribution of the data. This method is the bootstrap.

As stressed in Chapter 1, an important issue in Experimetrics is dependence, mainly with regard to the multiple observations per subject that are typically analysed. In this chapter, however, we sidestep such problems, by restricting attention to situations in which only one observation is available per subject (or two observations per subject in a within-subject test). The problem of treatment testing in the presence of dependence will be covered in the next chapter.

## 3.2 The Mechanics of Treatment Testing

Siegel & Castellan (1988) provide a useful summary of the mechanics of treatment testing. Here, we shall be somewhat brief.

A treatment test always has a null hypothesis and an alternative hypothesis. The null hypothesis is generally the hypothesis that there is no effect. The alternative hypothesis is that there is an effect. If the alternative hypothesis specifies the direction of the effect, it is a one-sided alternative and we conduct a one-tailed test. Otherwise 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. The first stage of the application of the test is to compute the test statistic which is a function of the data values. Then the test statistic is compared to the null distribution (i.e. the distribution that the statistic would in theory follow if the null hypothesis were true). If the test statistic falls in the rejection 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 $\alpha$, 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 more extreme than the one obtained. 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). The p-value represents the strength of evidence in favour of the alternative (i.e. evidence of an effect). The words used to represent "strength of evidence" are a matter of individual taste. Popular terminology is: 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.

Note that, in addition to considering whether there is an effect, and the strength of evidence of the effect, none of this is any use without also reporting the *direction* of the effect. As mentioned, 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. This is the value of prior beliefs in the form of economic theories.

## 3.3 Testing with Discrete Outcomes

### *3.3.1 The binomial test*

We will commence with what is perhaps the simplest of all tests.

Consider the following choice problem, where the two circles represent lotteries, and the areas within them represent probabilities of the stated outcomes. The left-hand lottery is the "safe" lottery and it pays \$5 with certainty. The right-hand lottery is the "risky lottery" and represents a 50:50 gamble involving the outcomes \$0 and \$10, as shown in Figure 3.1. Clearly, by choosing between these lotteries, a subject is conveying information about his or her attitude to risk. In particular, note that, since the two lotteries have the same expected value (\$5) (and assuming subjects obey expected utility theory), a risk-averse subject would choose S, a risk-seeking subject would choose R, and a risk-neutral subject would be indifferent between S and R. Accordingly, if all individuals were risk-neutral, we might expect 50% to choose the safe lottery, and 50% to choose risky. If, as is commonly found, individuals are predominantly risk averse, we would expect more than 50% to choose the safe lottery.

On this basis, the choices between the two lotteries of a sample of subjects may be used to conduct a test of the hypothesis of risk neutrality. The file **lottery_choice_sim** contains information on a sample of 30 subjects. The data set has 30 rows, one per subject, and one of the variables is y, consisting of ones and zeros representing choice of S and R respectively. The following command reveals that 21 of the 30 subjects chose S:

```
. tab y

          y |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          9       30.00       30.00
          1 |         21       70.00      100.00
------------+-----------------------------------
      Total |         30      100.00
```


Figure 3.1: A typical lottery choice problem

If the null hypothesis (risk neutrality) is true, the probability of any subject choosing S is 0.5. Therefore, if the null hypothesis is true, the probability of 21 or more of the 30 subjects choosing S is the following sum of binomial probabilities:

$$P(N_s \geq 21) = \sum_{n=21}^{n=30} \frac{30!}{n!(30-n)!}(0.5)^{30} = \underline{\underline{0.0213}} \tag{3.1}$$

Because this probability is smaller than 0.05, we may conclude that there is evidence that the null hypothesis is false, that is, that subjects are not risk-neutral (and, more precisely, that they are risk-averse).

The test that has just been performed is the binomial test, and the probability that has been computed is the p-value of the test. This test is used when the outcome is binary, and when the null hypothesis can be expressed simply in terms of the probability of one of the two outcomes.

The binomial test can be performed easily in STATA using the `bitest` command. Applying it to the present data set, we obtain the results:

```
. bitest y==0.5

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+--------------------------------------------------------------
           y |       30           21           15     0.50000     0.70000

Pr(k >= 21)                 = 0.021387  (one-sided test)
Pr(k <= 21)                 = 0.991938  (one-sided test)
Pr(k <= 9 or k >= 21)       = 0.042774  (two-sided test)
```

Three different p-values are provided. The first one is the same as the one computed above. Note that the test performed above is a one-sided test because the alternative hypothesis is that agents are risk averse, and hence we have a prior expectation that the proportion of safe choices will be greater than 0.5. If we had no such prior belief, we would use the two-sided test. Note that the p-value for the two-sided test is exactly two times that of the one-sided test. Note also that it is less than 0.05, meaning that even in the absence of the prior belief, we still have evidence in the sample that agents are not risk-neutral.

### 3.3.2 *Fisher's exact test*

Continuing with the same example, let us now assume that information on the subjects' gender (represented by the variable "male"; 1 if male, 0 if female) is also available. A cross-tabulation of gender and choice is obtained using the "tabulate" command:

```
. tab y male, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+
```

| y | male 0 | 1 | Total |
|---|---|---|---|
| 0 | 1<br>8.33 | 8<br>44.44 | 9<br>30.00 |
| 1 | 11<br>91.67 | 10<br>55.56 | 21<br>70.00 |
| Total | 12<br>100.00 | 18<br>100.00 | 30<br>100.00 |

We see that, of the 30 subjects, 12 are female and 18 are male. We also see that of the 12 female subjects, 11 (91.67%) chose S, while of the 18 males, 10 (55.56%) chose S. Note that these (column) percentages are shown as a result of using the "col" option. The difference in percentage choosing S appears to indicate that females are more risk averse than males. This difference is what needs to be tested for statistical significance.

For this purpose, we may use 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, for the given row totals and column totals. To consider how such a probability might be computed, let us consider the following tabulation:

| | Male=0 | Male=1 | Total |
|---|---|---|---|
| $Y=0$ | A | B | A+B |
| $Y=1$ | C | D | C+D |
| Total: | A+C | B+D | A+B+C+D |

The probability of obtaining this combination of numbers A, B, C, and D, for the given row and column totals, is given by:

$$P = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{A+B+C+D}{A+B}} = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{(A+B+C+D)!A!B!C!D!} \tag{3.2}$$

Applying (3.2) to the numbers in the cross-tabulation, we have:

$$P = \frac{9!21!12!18!}{30!1!8!11!10!} = 0.0367$$

Next, we need to ask what is the probability of obtaining a "more extreme" combination than the one above (i.e. one for which females appear to be even more risk-averse), for the given row and column totals. There is only one such combination:

| | Male=0 | Male=1 | Total |
|---|---|---|---|
| $Y=0$ | 0 | 9 | 9 |
| $Y=1$ | 12 | 9 | 21 |
| Total: | 12 | 18 | 30 |

Applying (3.2) to this combination, we obtain a probability of 0.0034. The probability that is required (i.e. the p-value for Fisher's exact test) is therefore:

$$0.0367 + 0.0034 = 0.0401$$

The test is performed in STATA using the "exact" option with the tabulate command:

```
. tab y male, col exact

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |          male
         y |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          8 |         9
           |      8.33      44.44 |     30.00
-----------+----------------------+----------
         1 |        11         10 |        21
           |     91.67      55.56 |     70.00
-----------+----------------------+----------
     Total |        12         18 |        30
           |    100.00     100.00 |    100.00

          Fisher's exact =                 0.049
  1-sided Fisher's exact =                 0.040
```

Two p-values are given. The "1-sided Fisher's exact" p-value is 0.040 and this agrees with the p-value computed above. Because it is less than 0.05, and because it is a one-sided test, we conclude that there is evidence that females are more risk-averse than males.

The other p-value is a two-tailed p-value. This is the p-value that we would use if we were testing for a gender effect in the absence of any prior belief over the direction of the effect. Computing this two-tailed p-value is slightly awkward. We need to add to the one-tailed p-value the probability of obtaining a more extreme outcome in the "other" direction. To determine whether an outcome is more extreme, we use the difference between the proportion of females choosing S and the proportion of males choosing S. For the data, this difference is $0.9167 - 0.5556 = 0.3611$.

We now need to consider various other possible outcomes. Consider:

| | Male=0 | Male=1 | Total |
|---|---|---|---|
| $Y = 0$ | 9 | 0 | 9 |
| $Y = 1$ | 3 | 18 | 21 |
| Total: | 12 | 18 | 30 |

For this outcome, the difference in proportions is $0.25 - 1 = -0.7500$. This is clearly more extreme than (and in the opposite direction to) 0.3611. The outcome

| | Male=0 | Male=1 | Total |
|---|---|---|---|
| $Y = 0$ | 8 | 1 | 9 |
| $Y = 1$ | 4 | 17 | 21 |
| Total: | 12 | 18 | 30 |

has a difference in proportions of $0.3333 - 0.9444 = -0.6111$. Again this is more extreme. The outcome

| | Male=0 | Male=1 | Total |
|---|---|---|---|
| $Y = 0$ | 7 | 2 | 9 |
| $Y = 1$ | 5 | 16 | 21 |
| Total: | 12 | 18 | 30 |

has a difference in proportions of $0.4166 - 0.8888 = -0.4722$. Yet again this is more extreme.

When the probability function (3.2) is applied to each of these three outcomes, we obtain (respectively) 0.00001, 0.00062, and 0.00847. The p-value for the two-sided test is therefore:

$$0.040 + 0.00001 + 0.00062 + 0.00847 = 0.0491.$$

This number is in agreement with the "Fishers exact" p-value given in the STATA output above. The fact that it is less than 0.05 indicates that there is evidence of a gender effect even in the absence of a prior belief over the direction of the effect.

### 3.3.3 The chi-squared test

The gender effect may alternatively be tested using the chi-squared test. This requires the `chi2` option with the `tab` command:

```
. tab y male, col chi2

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |          male
         y |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          8 |         9
           |      8.33      44.44 |     30.00
-----------+----------------------+----------
         1 |        11         10 |        21
           |     91.67      55.56 |     70.00
-----------+----------------------+----------
     Total |        12         18 |        30
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   4.4709   Pr = 0.034
```

We see that the p-value of this test is 0.034, slightly lower than that of Fisher's exact test, and provides further evidence of a gender difference in risk attitude.

Let us consider how the chi-squared statistic is computed. We need to ask what we would *expect* the numbers in the table above to be if the null hypothesis were true (i.e. if there really were no effect of gender). The answer is:

*Expected* frequencies:

| y | male 0 | male 1 | Total |
|---|---|---|---|
| 0 | 3.6<br>30.00 | 5.4<br>30.00 | 9<br>30.00 |
| 1 | 8.4<br>70.00 | 12.6<br>70.00 | 21<br>70.00 |
| Total | 12<br>100.00 | 18<br>100.00 | 30<br>100.00 |

We label the numbers in the first table O (for observed) and those in the second E (for expected), and we compute the following sum over the four cells in the table:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(1-3.6)^2}{3.6} + \frac{(11-8.4)^2}{8.4} + \frac{(8-5.4)^2}{5.4} + \frac{(10-12.6)^2}{12.6} = 4.4709$$

The test statistic has a $\chi^2$ distribution under the null hypothesis of no effect. What are the degrees of freedom? Here we need to ask how many of the four entries in the table are "free". Notice that if you fix the two entries in one of the rows, the two entries in the other row are determined (by the column totals); also, if you fix one of the entries in a row, the other is determined (by the row total). This means that there is only one free number in the table, and this is the degrees of freedom for the test. Generally, if the cross-tabulation has $m$ rows and $n$ columns, the degrees of freedom for the chi-squared test is $(m-1)(n-1)$.

Recall the critical values of the $\chi^2$ distribution:

| Degrees of freedom | 5% point of $\chi^2$ |
|---|---|
| 1 | 3.84 |
| 2 | 5.99 |
| 3 | 7.82 |
| 4 | 9.49 |
| 5 | 11.07 |
| 6 | 12.59 |

Here, we reject $H_0$ if $\chi^2 > 3.84$, which it is. As previously noted, this tells us that we have evidence of a difference between genders.

The p-value is the area to the right of the test statistic under the $\chi^2(1)$ distribution, i.e. p-value $= P(\chi^2(1) > 4.4709)$. The p-value of 0.034 being somewhat less than 0.05 is consistent with the test statistic being somewhat to the right of the 0.05 critical value. Note that this p-value may be computed exactly using =CHIDIST(4.4709,1) in Excel. This gives the same answer as STATA. of 0.034.

There is another way of carrying out the chi-squared test in STATA. Let us imagine that the complete data set is not available, and all that is available is the numbers in the cross tabulation. To carry out the chi-squared test with only this information, we would use the "immediate" STATA command `tabi`. Recall that an "immediate" command (which always ends with the letter i) is a command that obtains results not from the data stored in memory but from numbers typed as arguments. One such command seen previously in Chapter 2 is the `sampsi` command for finding the required sample size for a particular test. To reproduce the result of the last test using the command tabi, we use:

```
. tabi 1 8 \ 11 10, chi

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |         1          8 |         9
         2 |        11         10 |        21
-----------+----------------------+----------
     Total |        12         18 |        30

          Pearson chi2(1) =   4.4709   Pr = 0.034
```

### 3.3.4 *The chi-squared test on a real data set*

Various tests, starting with the chi-squared test, will be demonstrated in the context of the "hold-up" problem, using a data set from the experiment conducted by Ellingsen & Johannesson (2004). This data set is available in the STATA file **holdup**.

We shall start by describing the "holdup" experiment. There are two players: "seller" and "buyer". The game consists of three stages:

1. The seller is given 60 units, and has the opportunity to invest it in order to create the greater amount of 100 units. Note that this is a dichotomous choice: either he invests 60, or he does not invest at all.
2. If the seller has invested, the buyer proposes how to split the 100 units between the two players.
3. The seller chooses whether to accept the buyer's proposal, realising the proposed split, or to reject it. If the seller rejects, both parties receive zero, leaving the seller with a net loss of 60.

First of all, note that the holdup experiment is a combination of the trust game described in Section 2.5.3 and the ultimatum game described in Section 2.5.1. To be precise, the first two stages of the holdup experiment constitute a trust game, although it should be termed a "binary trust game" since the first mover's decision is either to invest or not to invest and is therefore binary. Stages 2 and 3 of the holdup experiment amount precisely to the ultimatum game.

Given conventional assumptions of self-interest, the unique subgame perfect equilibrium outcome is that the seller chooses not to invest. Why? At the last stage of the game, the seller should accept any proposal that gives him more than 0 units.

Hence there is no reason for the buyer to offer any more than 1 unit. Hence the seller expects to lose 59 units if he invests. So it is irrational for the seller to invest.

As usual, experimental findings cast serious doubt on the theory: about one-third of sellers choose to invest, and they often benefit from doing so.

Ellingsen & Johannesson (2004) focus on the effect of *communication* between buyer and seller. To this end, they consider three treatments:

Treatment 1 (T1): No communication except the actions themselves.

Treatment 2 (T2): Buyer can send a message to seller, before seller makes an investment decision (presumably, a message of the form "I am a principled person; if you reveal your trust in me by investing, I promise to reward you").

Treatment 3 (T3): Seller can send a message to buyer, along with the investment decision. (Presumably, the wording would be along the lines of "I am investing because I trust you, even though I do not know who you are. However, I also want you to know that I am not stupid; if I fail to benefit as a result of trusting you, I will make sure you receive zero".)

Clearly, the purpose of Treatment 2 is to assess the impact of *promises*, while Treatment 3 is to assess the impact of *threats*. Of course, neither promises nor threats alter the theoretical prediction of no investment. The issue is whether they have an impact on the actual decisions of either sellers or buyers.

*Research Question 1: What impact does communication have on the seller's decision of whether to invest?*

Here, we simply look at the proportion of investors for each treatment. The best way to look at these is in a cross-tabulation:

```
. tab invest treatment, col chi2

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

   sellers |
investment |
 decision: |
     1 if |
 invest; 0 |            treatment
   if not |         1          2          3 |     Total
----------+---------------------------------+----------
         0 |        26         14         12 |        52
           |     65.00      46.67      36.36 |     50.49
----------+---------------------------------+----------
         1 |        14         16         21 |        51
           |     35.00      53.33      63.64 |     49.51
----------+---------------------------------+----------
     Total |        40         30         33 |       103
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   6.1788   Pr = 0.046
```

Note that there are 103 pairs of subjects in the experiment. The overall proportion of investors is very close to 50%. However, there appear to be major differences

between treatments: the proportion of investors is highest in Treatment 3 (64%) and lowest in Treatment 1 (35%). This tells us that communication does have a favourable impact on the investment decision, especially communication in the form of the seller himself being in a position to make a threat (although we shall return later to the difference between the two types of communication).

To assess whether this difference between treatments is statistically significant, we conduct a chi-squared test, by including the `chi2` option in the STATA command above. We see that the p-value of this test is 0.046, indicating that there is evidence (although not strong evidence) that communication has an impact on the investment decision. An explanation of how the chi-squared test statistic is computed was provided in Section 3.3.3 above.

Note that the degrees of freedom for this test is 2. This is because the number of rows ($m$) is 2 while the number of columns ($n$) is 3, and the degrees of freedom is given by $(m-1)(n-1)$. As explained in Section 3.3.3, this essentially means that only two of the numbers in the table are "free"; the other four can always be deduced, with knowledge of the row and column totals.

Here, we reject $H_0$ if $\chi^2 > 5.99$, which it is. As previously noted, this tells us that we have evidence of a difference between treatments. The p-value of 0.046, being slightly less than 0.05, is consistent with the test statistic being slightly to the right of the 0.05 critical value.

We have established that communication does have an effect on the seller's decision to invest. However, we still need to establish whether one type of communication is more effective than the other.

*Research Question 2: Do the two different types of communication have differing effects on the seller's decision of whether to invest?*

This requires a chi-squared test again, but using only Treatments 2 and 3, i.e. using only two of the three columns of the contingency table:

```
. tab invest treatment if treatment!=1, chi2  col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

   sellers |
investment |
 decision: |
     1 if |
invest; 0 |        treatment
   if not |         2          3 |     Total
----------+----------------------+----------
         0 |        14         12 |        26
           |     46.67      36.36 |     41.27
----------+----------------------+----------
         1 |        16         21 |        37
           |     53.33      63.64 |     58.73
----------+----------------------+----------
     Total |        30         33 |        63
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   0.6882   Pr = 0.407
```

Note that the STATA command contains `if treatment!=1` which means "if treatment is not equal to 1", which results in the test being applied to a comparison of treatments 2 and 3.

While the percentage investing in T3 (63.64%) is somewhat higher than the percentage in T2 (53.33%), the chi-squared test reveals that there is no evidence of a difference between these. We therefore have no evidence that the two forms of communication differ in effectiveness with respect to the seller's decision.

## 3.4 Testing for Normality

In this section, we will continue to use the holdup data as an example, but we shall turn to the buyer's decision, given that the seller has invested. Unlike the seller's decision which is dichotomous, the buyer's decision is represented by an amount of money – the amount which he or she offers to return to the seller. The particular hypothesis in which we are interested in this section is normality of the data. Knowledge of whether the data is distributed normally is very useful in deciding between the different tests introduced in subsequent sections.

Since 51 of the sellers invested, we have 51 observations on the buyer's decision. The frequency distribution of these 51 offers is shown in Figure 3.2. The STATA syntax for obtaining this histogram is:

```
hist offer, disc freq normal xline(60) xlabel(0(10)100)
```

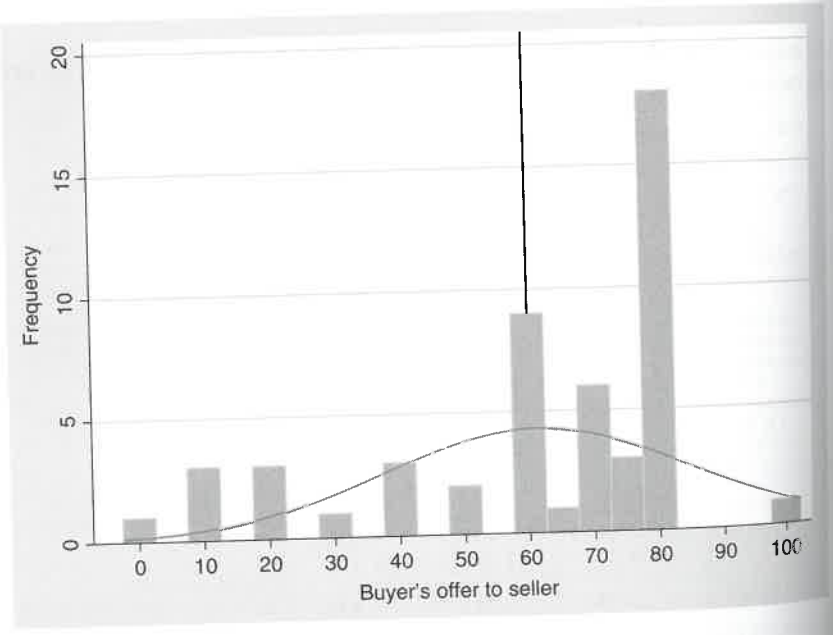



Figure 3.2: Frequency histogram of buyers' offers

This command has a number of options. The `disc` option ensures that there is a different bar for each discrete value of the data; `freq` causes frequency to be measured on the vertical axis, rather than density; `normal` causes a normal density to be superimposed on the histogram, with the same mean and standard deviation as the data; `xline(60)` causes a vertical line to appear at the offer of 60 (i.e. returning exactly the amount invested); `xlabel(0(10)100)` results in the x-axis covering the range 0 to 100 with tick-marks positioned every 10 units.

The modal buyer's offer is 80, indicating the tendency of the buyer to return the investment of 60 to the seller, in addition to splitting the profit of 40 evenly. However, some buyers offer amounts considerably below this, resulting in net losses for the seller.

The distribution does not appear to correspond closely to the normal curve. This is preliminary evidence that the data does not follow a normal distribution. A formal test of normality is the "skewness-kurtosis" test, obtained using the `sktest` command in STATA. Applying this test to the complete set of offer data, we obtain:

```
. sktest offer

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs    Pr(Skewness)    Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-----------------------------------------------------------------
       offer |     51       0.0021          0.4595          8.60          0.0136
```

The output actually contains three different test results, in the form of p-values. Pr(Skewness) is the p-value for the test of the hypothesis that skewness[1] equals zero (i.e. that the distribution is symmetric). The p-value of 0.0021 implies that symmetry is strongly rejected by the data. Pr(Kurtosis) is the p-value for the hypothesis of "normal kurtosis".[2] The third p-value represents the result of a joint test of skewness and kurtosis.

It is also useful to test for normality separately by treatment. We see that normality is rejected in Treatments 1 and 2, strongly in the latter.

```
. sktest offer if treatment==1

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs    Pr(Skewness)    Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-----------------------------------------------------------------
       offer |     14       0.5120          0.0088          6.52          0.0383

. sktest offer if treatment==2

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs    Pr(Skewness)    Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-----------------------------------------------------------------
```

[1] *Skewness* is measured by the third central moment of the distribution. Skewness is zero for a symmetric distribution. If skewness is positive, it is said that the distribution is "positively skewed" or "right-skewed", and the distribution is characterised by a long right-tail. Negative skewness (or left-skewness) is characterised by a long left-tail.

[2] *Kurtosis* is a measure of the fourth central moment of a distribution. For a standardised normal distribution, kurtosis is 3. If kurtosis is larger than 3, the distribution is said to be *leptokurtic* (fat-tailed); if less than 3, *platykurtic*.

```
       offer |     16       0.0003          0.0012         16.65          0.0002

. sktest offer if treatment==3

                    Skewness/Kurtosis tests for Normality
                                                                  ------- joint ------
    Variable |    Obs    Pr(Skewness)    Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------------------
       offer |     21       0.2886          0.6833         1.42          0.4918
```

Another test for normality is the Shapiro-Wilk test, conducted using the swilk command in STATA. When this test is applied to the complete sample, we reach the same conclusion as before: normality is strongly rejected.

```
. swilk offer

                    Shapiro-Wilk W test for normal data

    Variable |    Obs        W           V         z       Prob>z
-------------+-------------------------------------------------------
       offer |     51     0.87616      5.916     3.795     0.00007
```

## 3.5 Treatment Testing

### *3.5.1 Parametric tests of treatment effects*

In this section, we apply treatment tests to the buyer's decision in the holdup data. Because this decision is represented by a continuous variable (the amount which the buyer offers to return to the seller), a different set of tests are required from those introduced in the context of discrete outcomes in Section 3.3. Also, in Section 3.4, normality tests were applied to the buyer's offer, and evidence was found that the variable did not follow a normal distribution. This is important in choosing between the tests introduced in this section.

Since 51 of the sellers invested, we have 51 observations on the buyer's decision. The distribution of these 51 offers is shown in Figure 3.2. As noted in the previous section, the modal buyer's offer is 80, indicating the tendency of the buyer to return the investment of 60 to the seller, in addition to splitting the profit of 40 evenly. However, some buyers offer amounts considerably below this, resulting in net losses for the seller.

Again we are interested in any differences between treatments, and it is natural to compare "average" offers between the three treatments. The following table shows the mean offer by treatment:

```
. table treatment, contents(n offer mean offer)

------------------------------------
treatment |    N(offer)  mean(offer)
----------+-------------------------
        1 |          14      48.5714
        2 |          16           70
        3 |          21      63.3333
------------------------------------
```

Firstly, note that the cases in which no investment was made have been excluded from this table. This is because, when no investment is made, the offer is coded as a *missing value* (. in STATA). Please recognise the importance of this. A common mistake is to code missing observations as zeros – this would be very misleading here, since it would impose a severe downward bias on the mean offer.

Secondly, note that there appear to be differences between treatments, with communication tending to increase offers, this time with promises (T2) being more effective than threats (T3).

Again we need to consider whether these differences are statistically significant. There are three comparisons to be made:

*Research Question 3: What is the effect of seller communication (threats) on the buyers offer? (T3 vs T1)*

*Research Question 4: What is the effect of buyer communication (promises) on the buyers offer? (T2 vs T1)*

*Research Question 5: Do the two forms of communication differ in their impact on buyers offers? (T3 vs T2)*

There are a number of possible ways to address these questions.

#### *3.5.1.1 The independent samples t-test (or two-sample t-test)*

We are comparing two samples. Let us assume that the first sample comes from a population with mean $\mu_1$ and standard deviation $\sigma_1$, while the second sample comes from a population with mean $\mu_2$ and standard deviation $\sigma_2$.

We wish to use the information in the two samples to test the null hypothesis $H_0$: $\mu_1 = \mu_2$ against the alternative $H_1$: $\mu_1 \neq \mu_2$. The information in the two samples is in the form of sample sizes $n_1$ and $n_2$, sample means $\bar{x}_1$ and $\bar{x}_2$, and sample standard deviations $s_1$ and $s_2$.

The independent samples t-test statistic is based straightforwardly on the difference between the two sample means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.3}$$

where $s_p$ (the pooled standard deviation) is just a weighted average of the two individual sample standard deviations:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} \tag{3.4}$$

The pooled standard deviation ($s_p$) is used when it is assumed that the two populations have the same variance, i.e. that $\sigma_1 = \sigma_2$. If there are reasons for not assuming this, there is also an "unequal variances" version of the test.

Given that certain assumptions are met, the t-statistic presented in (3.3) has a $t(n_1 + n_2 - 2)$ distribution under the null hypothesis.

Let us apply the test to Research Question 3 (seller communication):

```
. ttest offer if treatment!=2, by(treatment)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    48.57143    8.619371   32.25073    29.95041    67.19245
       3 |      21    63.33333    4.230464   19.38642    54.50874    72.15793
---------+--------------------------------------------------------------------
combined |      35    57.42857    4.383753   25.93463    48.51971    66.33743
---------+--------------------------------------------------------------------
    diff |           -14.7619    8.711774               -32.48614    2.962333
------------------------------------------------------------------------------
    diff = mean(1) - mean(3)                                      t =  -1.6945
Ho: diff = 0                     degrees of freedom =       33

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0498         Pr(|T| > |t|) = 0.0996          Pr(T > t) = 0.9502
```

The test statistic is $-1.6945$, which is compared with the t(33) distribution. This is done for us. Various p-values are shown in the final row of the results. The "two-tailed p-value" is seen to be 0.0996. This represents mild evidence of a difference between the two treatments (1 and 3).

In this situation, we have a prior belief about the direction of the effect: we expect seller communication to have a positive effect on buyer offer. For this reason we only reject the null if the statistic is in the lower tail of the distribution, and we divide the p-value by 2, giving 0.0498. This is the p-value shown on the left. Note that having a prior belief allows us to interpret the evidence as being stronger, and we are able to upgrade the evidence from "mild evidence" to simply "evidence".

As noted above, there is a version of the test that can be used if the two population variances are not assumed to be equal. This just requires the `unequal` option:

```
. ttest offer if treatment!=2, by(treatment) unequal

Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    48.57143    8.619371   32.25073    29.95041    67.19245
       3 |      21    63.33333    4.230464   19.38642    54.50874    72.15793
---------+--------------------------------------------------------------------
combined |      35    57.42857    4.383753   25.93463    48.51971    66.33743
---------+--------------------------------------------------------------------
    diff |           -14.7619    9.601583               -34.83782    5.314008
------------------------------------------------------------------------------
    diff = mean(1) - mean(3)                                      t =  -1.5374
Ho: diff = 0     Satterthwaite's degrees of freedom =     19.79

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0702         Pr(|T| > |t|) = 0.1404          Pr(T > t) = 0.9298
```

Note that the evidence is downgraded to "mild" as a consequence of not assuming equal variances. This is a common experience in hypothesis testing; the less that we can assume before conducting a test, the weaker the evidence is likely to be.

If you really wish to know whether you can assume equal variances, conduct a variance-ratio test, as follows:

```
. sdtest offer if treatment!=2, by(treatment)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    48.57143    8.619371    32.25073    29.95041    67.19245
       3 |      21    63.33333    4.230464    19.38642    54.50874    72.15793
---------+--------------------------------------------------------------------
combined |      35    57.42857    4.383753    25.93463    48.51971    66.33743
------------------------------------------------------------------------------
    ratio = sd(1) / sd(3)                                     f =    2.7675
Ho: ratio = 1                         degrees of freedom =    13, 20

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
 Pr(F < f) = 0.9801          2*Pr(F > f) = 0.0398          Pr(F > f) = 0.0199
```

This unfortunately tells us that there is evidence of a difference in variances between treatments 1 and 3.

The two-sample t-test is applied to the other two research questions. Results from all three (ignoring the problem of unequal variances) are shown below:

| | | Two-tailed p-value (equal variances) |
|---|---|---|
| Q3 | No comm. (T1) vs seller comm. (T3) | 0.0996 |
| Q4 | No comm. (T1) vs buyer comm. (T2) | 0.0250 |
| Q5 | Seller comm. (T3) vs buyer comm. (T2) | 0.2673 |

Two-tailed p-values from two-sample t-tests for treatment effects on buyer's offer.

As remarked above, the two-sample t-test relies on quite strong assumptions about the data. Most importantly, unless the two samples happen to be "large", it is required that the two populations are normally distributed. In Section 3.4 we found strong evidence that the buyer's offer is not normally distributed. This result, together with the fact that the numbers of observations in each treatment are considerably lower than the 30 required for the central limit theorem to apply, leads us to doubt the validity of the tests carried out in this sub-section.

### 3.5.2 *Non-parametric tests of treatment effects: the Mann-Whitney test*

A test for comparing two samples which does not rely on any distributional assumptions (such as normality of the data) is the Mann-Whitney U test. Because no such assumptions are made, it is classed as a non-parametric test.

To carry out the 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.

The test is carried out in STATA using the `ranksum` command. The following compares T1 and T3, and is therefore a test of Research Question 3.

```
. ranksum offer if treatment!=2, by(treatment)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

   treatment |       obs    rank sum    expected
-------------+---------------------------------
           1 |        14       220.5         252
           3 |        21       409.5         378
-------------+---------------------------------
    combined |        35         630         630

unadjusted variance        882.00
adjustment for ties        -26.56
                         ----------
adjusted variance          855.44

Ho: offer(treatm~t==1) = offer(treatm~t==3)
             z =  -1.077
    Prob > |z| =   0.2815
```

The p-value of 0.2815 indicates that there is no evidence of a difference in offers between T1 and T3.

The Mann-Whitney test may be applied to all of Research Questions 3–5, and the results, in terms of p-values, are presented (and compared with the corresponding results from t-tests) in the following table:

| | | Two-sample t-test | Mann-Whitney test |
|---|---|---|---|
| Q3 | T1 vs T3 | 0.0996 | 0.2815 |
| Q4 | T1 vs T2 | 0.0250 | 0.0886 |
| Q4 | T3 vs T2 | 0.2673 | 0.1765 |

Two-tailed p-values from two-sample t-tests and Mann-Whitney tests

Using the Mann-Whitney test, we do not find any significant differences, with the exception of Q4, for which we find only mild evidence of a difference. Intuitively, we expect non-parametric tests such as this one to indicate weaker evidence of an effect than the corresponding parametric test, simply because fewer assumptions are being made. Comparisons of p-values seen in this table with those in the previous table are broadly consistent with this expected pattern.

### 3.5.3 *The bootstrap*

We have introduced the Mann-Whitney test as a non-parametric analogue to the two-sample t-test, 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 & 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. The

technique was applied by Ellingsen & Johannesson (2004) to their holdup data. We will attempt to reproduce their results below.

The bootstrap procedure consists of 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, $\hat{t}_j^*$, $j = 1, \ldots, B$.
3. Compute the standard deviation $s_B$ of the bootstrap test statistics $\hat{t}_j^*$, j=1,. . . , $B$.
4. Obtain the new test-statistic $z_B = \hat{t}/s_B$.
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 $\alpha$ is the chosen test size. Since $\alpha$ 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. This recommendation is followed here.

The following command applies the bootstrap two-sample t-test to Research Question 3 (T3 vs. T1):

```
bootstrap t=r(t), rep(999) nodrop : ///
ttest offer if treatment!=2, by(treatment)

(running ttest on estimation sample)

Bootstrap replications (999)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................   50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................

Bootstrap results                                 Number of obs      =       103
                                                  Replications       =       999

      command:  ttest offer if treatment!=2, by(treatment)
            t:  r(t)

-------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
           t |  -1.694477   1.144873    -1.48   0.139    -3.938386    .5494314
-------------------------------------------------------------------------------
```

Since the `bootstrap` command is too long to fit onto one line, we have spread it over two lines with the continuation marker "///". When the command is run, a sequence of dots appears on the screen, one for each bootstrap sample, so that the user is aware of the progress of the procedure.

Another useful thing to do when learning this technique is to add the `saving` option to the `bootstrap` command. In the next command, we use this option, along with the higher number of bootstrap samples, 9,999. Since we do not want 9,999 dots to appear on the screen, we also use the option `nodots`.

```
bootstrap t=r(t), nodots rep(9999) nodrop saving("hello.dta", replace) : ///
ttest offer if treatment!=2, by(treatment)

Bootstrap results                                 Number of obs      =       103
                                                  Replications       =      9999

      command:  ttest offer if treatment!=2, by(treatment)
            t:  r(t)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -1.694477   1.1553     -1.47   0.142    -3.958824    .5698692
------------------------------------------------------------------------------
```

When the `saving` option is used, the bootstrap test statistics $\hat{t}_j^*$, j=1,...,9999 are stored in the new data set "hello". If we read the contents of this file, we can then investigate the distribution of the 9,999 bootstrap test statistics:

```
. summ t

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           t |      9999   -1.718972      1.1553  -8.077973   2.979723
```

Note in particular that the standard deviation of these 9,999 numbers is the "bootstrap standard error" that appears in the results table from the bootstrap command shown above. The distribution of bootstrap test statistics is also shown as a histogram in Figure 3.3. We see that the distribution is bell-shaped and symmetric, with the centre very close to the "actual" t-statistic for the test, which was −1.6945.

We can now add the bootstrap column to the table of test results for Research Questions 3–5. This is done in the following table. These results are similar to those appearing in Table 2 of Ellingsen & Johannesson (2004), although they are not identical and this is an expected consequence of the randomness implicit in the bootstrap procedure. We agree with Ellingsen & Johannesson (2004) that the comparison of T1 and T2 (buyer communication) is the only one that shows a significant effect.

| | | Two-sample t-test | Mann-Whitney test | Bootstrap |
|---|---|---|---|---|
| Q3 | T1 vs T3 | 0.0996 | 0.2815 | 0.140 |
| Q4 | T1 vs T2 | 0.0250 | 0.0886 | 0.045 |
| Q5 | T3 vs T2 | 0.2673 | 0.1765 | 0.289 |

Two-tailed p-values from: two-sample t-tests; Mann-Whitney tests; bootstrap tests

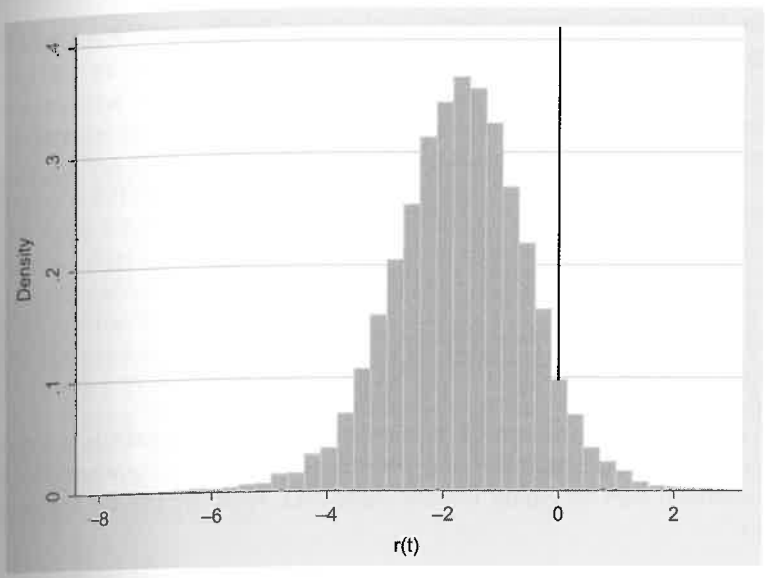



Figure 3.3: The distribution of the 9,999 bootstrap test statistics

### *3.5.4 Tests comparing entire distributions*

In this section, we consider data from the ultimatum game and the dictator game. The reader is referred back to Section 2.5 for an explanation of these games, a discussion of their theoretical predictions, and a summary of empirical evidence relating to them.

Forsythe et al. (1994) set out to test whether fairness alone can explain proposers' willingness to give to responders in these settings. They do this by subjecting one group to an ultimatum game, and the other group to a dictator game. The basic idea is that, if giving is the same in both, then fairness is the only explanation for giving. If giving is greater in the ultimatum game, then other factors must be influencing the decision (e.g. fear of the offer being rejected).

The tests that they favour are based on a comparison of the entire distributions of proposals under the two treatments, rather than a comparison of a particular characteristic of the distribution such as mean or variance. This is because (Forsythe et al., 1994, p. 351):

> conventional theory predicts that proposals will be concentrated at a single point... Since theory does not predict a distribution of proposals, it provides no guidance about which functionals of the distribution should be tested. Invariance of the entire distribution has the appealing property of implying that all functionals are invariant.

Tests that make comparisons of entire distributions include: the Kolmogorov-Smirnov test; the Epps-Singleton test; the Cramer-von Mises test; the Anderson-Darling test. We will demonstrate the first two of these on the data of Forsythe et al. (1994). The data is contained in the file **forsythe**.

The Kolmogorov-Smirnov test is implemented by the command `ksmirnov` in STATA. Results are as follows:

```
. ksmirnov y, by(dic_ult)

Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D        P-value  Corrected
 ------------------------------------------------
 1:                  0.3516     0.000
 2:                 -0.0110     0.989
 Combined K-S:       0.3516     0.000      0.000

Note: ties exist in combined dataset;
      there are 11 unique values out of 182 observations.
```

In order to understand how the Kolmogorov-Smirnoff test statistic is computed, a very useful graph is a "cdfplot" (this is a user-written STATA command that needs to be installed; start by typing `findit cdfplot`). Here, we use the command as follows:

```
cdfplot y, by(dic_ult)
```

and the result is shown in Figure 3.4. The higher of the two lines in the plot is the cumulative distribution function (cdf) of giving in the dictator game; the lower is the same for the ultimatum game. That the dictator game cdf is higher is consistent with dictator game offers typically being lower than ultimatum game offers.

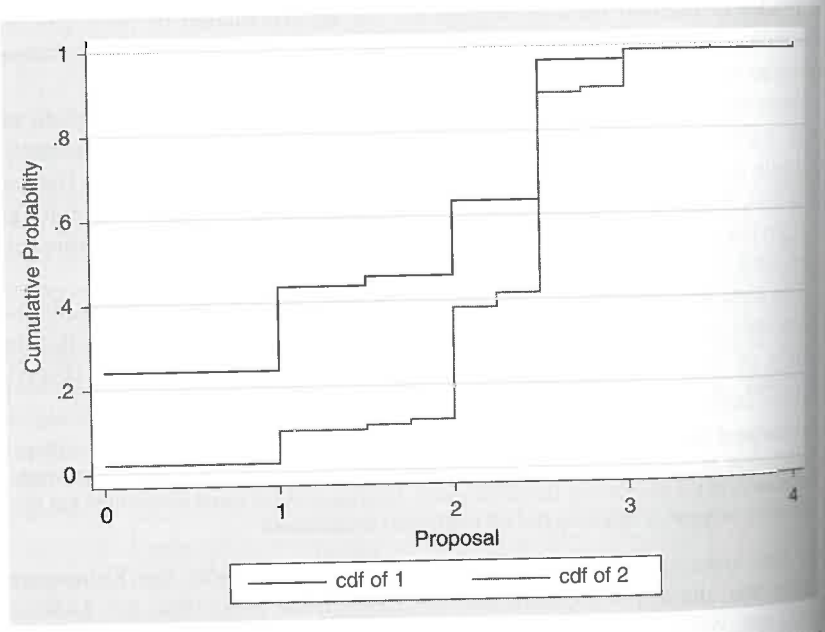



Figure 3.4: cdfs of giving in the ultimatum game (lower line) and dictator game (higher line)

The Kolmogorov-Smirnov test statistic is computed as the largest vertical distance between the two cdfs, and we see from the STATA output, and also from the plot, that this maximal distance (occurring for proposals between 1.5 and 1.75) is 0.3516. This difference is then compared to a null distribution in order to obtain the p-value for the test, which is 0.000. This p-value indicates that there is overwhelming evidence of a difference between the two distributions.

The Epps-Singleton test (Epps & Singleton, 1986) does not compare the two distributions directly, but instead compares the empirical characteristic functions. 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). This is another command that needs to be installed; start by typing `findit ecftest`.

Here is the output from applying the Epps-Singleton test to the comparison of dictator and ultimatum giving.

```
. escftest y, group(dic_ult)

Epps-Singleton Two-Sample Empirical Characteristic Function test

Sample sizes: dic_ult = 1            91
              dic_ult = 2            91
              total                 182
t1                                0.400
t2                                0.800

Critical value for W2 at 10%       7.779
                         5%        9.488
                         1%       13.277
Test statistic W2                 35.624

Ho: distributions are identical
P-value                      0.00000
```

We have applied both the Kolmogorov-Smirnov test and the Epps-Singleton test to the problem of comparing the distributions of giving in the ultimatum and dicatator games. Both tests result in a p-value of 0.000, amounting to overwhelming evidence of a difference between the two distributions. The interpretation of this result is that fairness is not the only consideration that enters the decision of how much to give.

## 3.6 Testing for Gender Effects

In this section, we show how treatment tests can be used to test for a gender effect. This is done by simply treating gender as the "treatment", although it is clearly not a typical treatment since it is not assigned by the experimenter. This will be done in the context of the ultimatum game, which was explained in Section 2.5. As in the last section, we are interested in the proposer's decision.

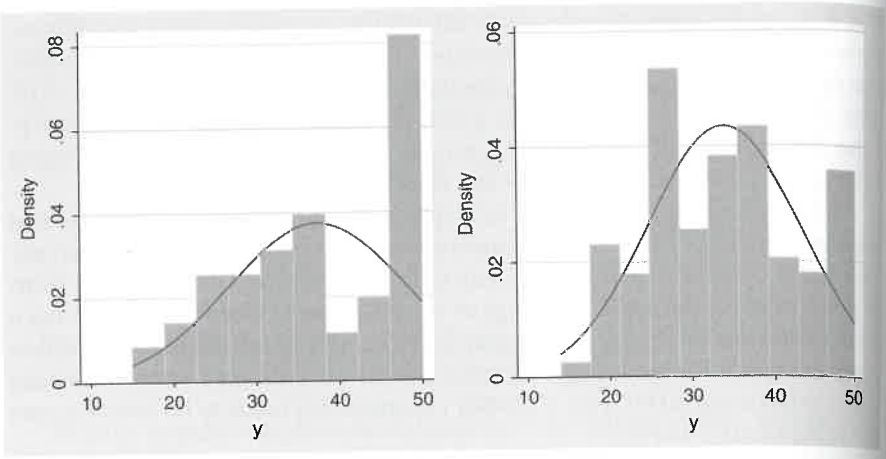



Figure 3.5: Proposers offers: females (left); male(right)

The file **ug_sim** contains (simulated) data from 200 subjects who participated in an ultimatum game, in which the size of the pie is 100 units. Each subject plays twice, once as proposer, and once as responder, with a different opponent each time. The variables are:

| | |
|---|---|
| i: | Proposer ID |
| j: | Responder ID |
| male_i: | 1 if proposer is male; 0 otherwise |
| male_j: | 1 if responder is male; 0 otherwise |
| y: | Proposer's offer |
| d: | Responder's decision: 1 if accept; 0 if reject |

A large amount of research has been done on gender effects in the ultimatum game. A good paper to start with is Eckel & Grossman (2001). Here, we will look for gender differences in proposers' offers. The distributions of proposers' offers are shown separately by gender in Figure 3.5. As expected, offers are distributed between zero and 50 (one half of the pie), and there is an accumulation of offers at 50 for both genders. Normal densities are superimposed. Neither distribution appears close to the normal.

Next, we test formally for normality in the distribution of proposer offers, for each gender separately, using one of the tests for normality introduced in Section 3.4.

```
. sktest y if male_i==0

                    Skewness/Kurtosis tests for Normality
                                                          ------- joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+------------------------------------------------------------------
           y |    91       0.3391         0.0000        21.61          0.0000
```

```
. sktest y if male_i==1

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+------------------------------------------------------------------
           y |   109       0.3899          0.0135         6.42          0.0403
```

Normality is rejected for both genders. For females, rejection is particularly strong, and this is a consequence of females having a larger accumulation (than males) of observations at the equitable allocation of 50 (see Figure 3.5).

The next thing we might wish to do is to test for equal variances between the two samples. The results of this test are as follows:

```
. sdtest y, by(male_i)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      91    37.37363    1.115618    10.64231    35.15726      39.59
       1 |     109    33.86239    .8779076    9.165624    32.12222    35.60255
---------+--------------------------------------------------------------------
combined |     200       35.46    .7067101     9.99439      34.0664    36.8536
------------------------------------------------------------------------------
    ratio = sd(0) / sd(1)                                     f =    1.3482
Ho: ratio = 1                             degrees of freedom =  90, 108

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
 Pr(F < f) = 0.9314          2*Pr(F > f) = 0.1372          Pr(F > f) = 0.0686
```

This test shows mild evidence that the two variances are different. We might wish to allow for these difference in variance in the next test, the independent samples t-test for a gender difference in proposer offers:

```
. ttest y, by(male_i)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      91    37.37363    1.115618    10.64231    35.15726      39.59
       1 |     109    33.86239    .8779076    9.165624    32.12222    35.60255
---------+--------------------------------------------------------------------
combined |     200       35.46    .7067101     9.99439      34.0664    36.8536
---------+--------------------------------------------------------------------
    diff |            3.511241    1.400706                .7490252    6.273457
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.5068
Ho: diff = 0                                  degrees of freedom =      198

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9935         Pr(|T| > |t|) = 0.0130          Pr(T > t) = 0.0065

. ttest y, by(male_i) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      91    37.37363    1.115618    10.64231    35.15726      39.59
       1 |     109    33.86239    .8779076    9.165624    32.12222    35.60255
```

```
---------+--------------------------------------------------------------------
combined |      200        35.46     .7067101     9.99439      34.0664      36.8536
---------+--------------------------------------------------------------------
    diff |            3.511241    1.419621                     .7098767     6.312605
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.4734
Ho: diff = 0                     Satterthwaite's degrees of freedom =  178.831

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9928         Pr(|T| > |t|) = 0.0143          Pr(T > t) = 0.0072
```

Whether or not we assume equal variances, there is strong evidence of a gender difference, with females offering more than males on average. The evidence is slightly stronger (indicated by the slightly smaller p-value) when equal variances are assumed.

Despite the rejection of normality, this is a situation in which the result of the independent samples t-test may be relied upon, by virtue of the large sample sizes. However, for good measure, we shall also conduct a non-parametric test for a gender effect.

```
. ranksum y, by(male_i)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      male_i |      obs    rank sum    expected
-------------+---------------------------------
           0 |       91     10122.5      9145.5
           1 |      109      9977.5     10954.5
-------------+---------------------------------
    combined |      200       20100       20100

unadjusted variance   166143.25
adjustment for ties    -1147.79
                     ----------
adjusted variance     164995.46

Ho: y(male_i==0) = y(male_i==1)
             z =   2.405
    Prob > |z| =   0.0162
```

Again we see evidence of a gender difference, although the effect is not strong this time, since the p-value is greater than 0.01. A general conclusion from this sequence of tests is that the more that can be assumed about the data (i.e. the more "parametric" the test) the stronger the test result tends to be (in terms of the closeness to zero of the p-value).

We could, of course, apply other tests to this problem, such as the tests of equality of entire distributions introduced in Section 3.5.4. We would expect to reach a similar conclusion. These tests are left to the reader.

## 3.7 Within-subject Tests

Within-subject tests are used to test the effect of a treatment in a situation in which each subject is observed both before and after the treatment. This is in contrast to

the between-subject tests that have been considered up until now in this chapter, in which one group of subjects are exposed to a treatment, while a different group is not exposed. From a theoretical point of view, within-subject tests are preferred to between-subject tests; they have more statistical power. However, there are various reasons why within-subject tests are not favoured by experimental economists. The issue of "order effects" is much discussed (see for example Harrison et al., 2005; Holt & Laury, 2002); an order effect is present if the result of the test depends on the order in which the control and the treatment are administered. More generally, there are concerns that the experience of one treatment impacts on behaviour in the treatment that follows.

There are however some instances in experimental economics in which within-subject tests are the most natural approach.

The test that is most appropriate to many within-subject settings arising in experimental economics is the McNemar change test (see Siegel & Castellan, 1988). This is because the two decisions are usually binary, and we are simply interested in the subjects who "switch" from one choice to the other, and in the direction in which they are switching. The Conlisk (1989) test is an alternative test that is applicable in this setting. If the outcome observed on the two occasions has a continuous distribution, it is appropriate to use the paired-comparison t-test, or, if a non-parametric test is preferred, the Wilcoxon signed ranks test.

### *3.7.1 The Allais paradox*

The Allais paradox (Allais, 1953) is perhaps the most well-known contradiction of expected utility (EU) theory. It is normally tested using within-subject tests of the type introduced above.

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, A*B*, AB*, 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. To test this null, we apply the chi-squared test introduced in Section 3.3.3 to these two groups. The test statistic is:

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{\left[n(AB^*) - \frac{n(AB^*)+n(A^*B)}{2}\right]^2}{\frac{n(AB^*)+n(A^*B)}{2}} + \frac{\left[n(A^*B) - \frac{n(AB^*)+n(A^*B)}{2}\right]^2}{\frac{n(AB^*)+n(A^*B)}{2}} \tag{3.5}$$

Expanding and simplifying (3.5), we obtain:

$$\chi^2 = \frac{\left[n(AB^*) - n(A^*B)\right]^2}{n(AB^*) + n(A^*B)} \tag{3.6}$$

The distribution of (3.6) is $\chi^2(1)$ under the null hypothesis of no Allais behaviour.[3]

Conlisk (1989) presented the two choice problems to 236 subjects. The numbers providing each response combination are given in the following table:

| | B | $B^*$ |
|---|---|---|
| A | 18 | 103 |
| $A^*$ | 16 | 99 |

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

It is the off-diagonal entries in the table on which we focus, and we see that the number of subjects answering $AB^*$ is considerably greater than the number answering $A^*B$. To test the difference statistically, we apply the McNemar test (3.6):

$$\chi^2 = \frac{[103 - 16]^2}{103 + 16} = 63.6$$

[3] In Section 3.3.3 it was explained that the null-distribution of the chi-squared test based on a $m \times n$ cross-tabulation is $\chi^2$ with $(m-1)(n-1)$ degrees of freedom. An exception arises when either $m = 1$ or $n = 1$, that is, when the cross-tabulation consists of only one row or one column. If $m = 1, df = n - 1$ and if $n = 1, df = m - 1$. In the present context, the cross-tabulation is essentially $2 \times 1$, and hence the degrees of freedom is one.

Since the null distribution is $\chi^2(1)$, any value of this test statistic greater than 3.84 would constitute evidence of Allais behaviour, and any value greater than 6.63 constitutes strong evidence. The statistic of 63.6 therefore represents strong evidence of Allais behaviour.

One further point about the McNemar test is that when the numbers appearing in the formula 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). The formula for the test statistic including the continuity correction is given by:

$$\chi^2 = \frac{\left[|n(AB^*) - n(A^*B)| - 1\right]^2}{n(AB^*) + n(A^*B)} \tag{3.7}$$

In the present case, the test statistic changes from 63.6 to 62.15 when the continuity correction (3.7) is applied.

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 the following formula:

$$Z = \frac{\sqrt{N-1}\left(S - \frac{1}{2}\right)}{\sqrt{\frac{1}{4V} - \left(S - \frac{1}{2}\right)^2}} \tag{3.8}$$

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

$$V = \frac{n(AB^*) + n(A^*B)}{N} \tag{3.9}$$

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

$$S = \frac{n(AB^*)}{n(AB^*) + n(A^*B)} \tag{3.10}$$

The test statistic (3.8) has a standard normal distribution under the null hypothesis of no Allais behaviour. A value in the upper tail of the standard normal distribution provides evidence that the proportion $S$ is significantly greater than one half, that is, evidence of Allais behaviour.

Applying the Conlisk test to the data in the table, we obtain $V = 0.504$, $S = 0.866$, and

$$Z = \frac{\sqrt{236-1}\left(0.866 - \frac{1}{2}\right)}{\sqrt{\frac{1}{4\times 0.504} - \left(0.866 - \frac{1}{2}\right)^2}} = \underline{\underline{9.32}}$$

This test statistic is certainly in the upper tail of the standard normal distribution, again providing evidence of Allais behaviour in this sample.

### *3.7.2 Preference reversals*

"Preference reversal" (PR) is a term normally used to refer to 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 (the "$-bet") when asked to value them (that is, to provide their certainty equivalent) separately. The phenomenon was apparently discovered by Lichtenstein & Slovic (1971), and later introduced to the economics literature by Grether & Plott (1979).

For a particular example, let us look to the study of Tversky et al. (1990). The very first pair of lotteries they consider (in their Table 1: study 1, set 1, triple 1) is:

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

The number of subjects they presented with this pair of lotteries was 179. From the information in their Table 2, we may deduce the following.

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

Results from Tversky et al. (1990), study 1, set 1, triple 1.

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". If the number of subjects making standard reversals is significantly greater than the number making non-standard reversals, we may conclude there is evidence of the PR phenomenon.

This is clearly a situation in which "within-subject" tests are essential. It is obvious that for a PR to be observable, it is necessary for the same subject to be "observed" twice: once making the choice, and once reporting their two valuations. The tests which are appropriate are the same as those used in Section 3.7.1 for testing Allais behaviour.

Applying the McNemar test to the numbers in the above table, we obtain:

$$\chi^2 = \frac{[106-4]^2}{106+4} = 94.58$$

and, being $\chi^2(1)$ under the null, this amounts to strong evidence of the PR phenomenon.

Applying the Conlisk test, we obtain $V = 0.614$, $S = 0.963$, and

$$Z = \frac{\sqrt{179-1}\left(0.963-\frac{1}{2}\right)}{\sqrt{\frac{1}{4\times 0.614}-\left(0.963-\frac{1}{2}\right)^2}} = \underline{\underline{14.07}}$$

Being standard normal under the null, this positive test statistic amounts to strong evidence of the PR phenomenon.

### *3.7.3 Continuous outcome*

All of the within-subject tests considered so far have been in the context of binary outcomes. We now turn to the situation in which subjects are again observed both without and with a treatment, but the outcome is a continuous variable.

One such situation arises when investigating the impact of a "take treatment" in a dictator game experiment. A "take game" is a dictator game in which dictators are allowed to take money away from the recipient, that is, to "give" less than zero. Bardsley (2008) and List (2007) find that dicator game giving is lower when this opportunity to take is introduced. Both of those studies use between-subject tests. However, an obvious alternative approach is a within-subject design in which subjects play two dictator games in succession: the first a "give only" game; the second a "give or take" game. The amount given may then be compared between the two treatments in order to test the effect of the "take treatment". To our knowledge, Chlaß & Moffatt (2012) are the only authors to have adopted the within-subject approach in this particular setting.

Here, we assume the following design. Subjects each play two dictator games with a pie of size 10 units. In the first game, they are asked how much, if any, of the pie they would like to give to the recipient. In the second game, they are again asked how much they would like to give to the recipient, but their opportunity set is extended such that they are allowed to "give" a negative amount up to 10 units; that is, they are allowed to *take* up to 10 units from the recipient. After they have made both decisions, one of the two games is selected by a random device, and the payoffs are implemented in accordance with the decision made in the selected game.

The file **give_take_sim** contains simulated data from 50 subjects. The variables are:

i: subject id
y1: giving in give-only game
y2: giving in the give-or-take game ($y2 < 0$ if amount is taken from recipient)

It is useful to start by plotting the two giving variables against each other. We use the following scatter command:

```
scatter y2 y1, msize(1)  jitter(2) yline(0) xlabel(0(1)5) ylabel(-5(1)5)
```

The scatterplot is shown in Figure 3.6. The `jitter` option is useful in this situation because it applies small random perturbations to the position of each point in the scatter, making it possible to see in which locations there are large accumulations of points. We see that a significant number of subjects lie on the 45 degree line, implying that the amount they give is the same in both treatments. However, some subjects lie below the 45 degree line, implying that when there is an option to take,

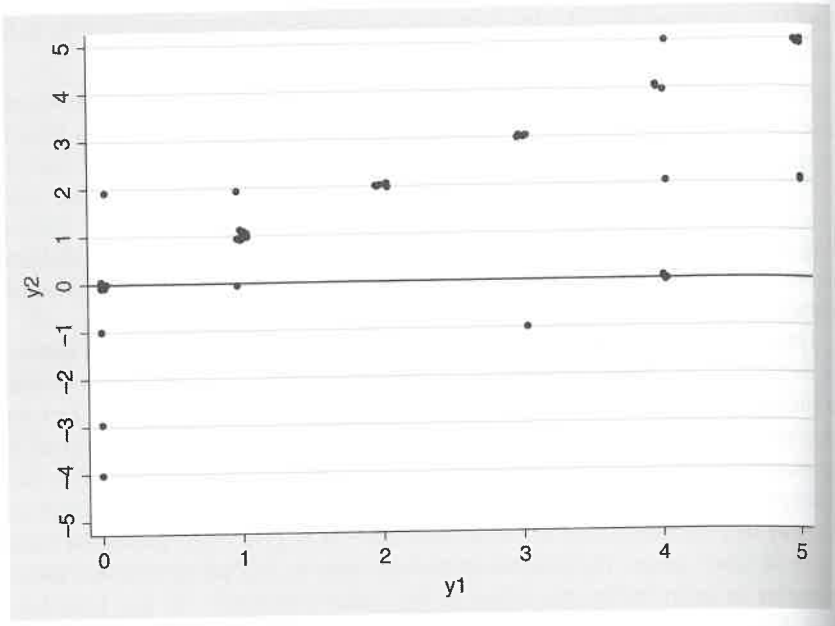



Figure 3.6: Giving in a give-or-take game against giving in a give-only game

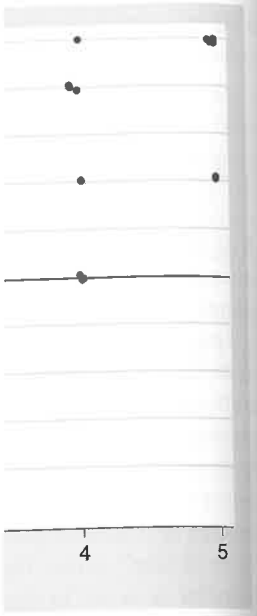
they give less, and sometimes "give" negative amounts. Very few subjects are above the 45 degree line.

Readers may be confused that giving is being described as a "continuous" outcome while it is evident from Figure 3.6 that giving in fact takes only a small number of discrete values. The important point here is that the variable "giving" is, in theoretical terms, a continuous variable. The fact that the observed variable is discrete is simply a consequence of the manner in which the variable has been measured – namely, inducing subjects to select a whole number. This is, of course, a feature of all measurement systems: all continuous variables must be measured at some level of rounding. It is possible to deal with rounding econometrically, by estimating the interval regression model, which will be covered in Chapter 6. However, it can be verified that in situations like the current one, applying such a model yields results very similar to those obtained by treating the outcome variable as a continuous variable, as done here.

To test formally for a treatment effect, we may, as usual, choose between a parametric and a non-parametric test. The parametric test is the paired comparisons t-test. This test computes the difference in giving between treatments for each observation, and then applies the t-test to test whether these differences have mean zero. The results are:

```
. ttest y2=y1

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      y2 |      50        1.56    .2971429    2.101117    .9628691    2.157131
      y1 |      50        2.18    .2613778     1.84822    1.654742    2.705258
```




g in a give-only game

Very few subjects are above

ibed as a "continuous" out-
in fact takes only a small
; that the variable "giving"
t that the observed variable
which the variable has been
number. This is, of course,
variables must be measured
unding econometrically, by
covered in Chapter 6. How-
one, applying such a model
g the outcome variable as a

as usual, choose between a
est is the paired comparisons
en treatments for each obser-
differences have mean zero.

| v. | [95% Conf. | Interval] |
|---|---|---|
| .7 | .9628691 | 2.157131 |
| 2 | 1.654742 | 2.705258 |



```
---------+--------------------------------------------------------------------
    diff |       50        -.62    .2038206     1.44123   -1.029593   -.2104071
---------+--------------------------------------------------------------------
     mean(diff) = mean(y2 - y1)                                     t =  -3.0419
Ho: mean(diff) = 0                                     degrees of freedom =       49

Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
Pr(T < t) = 0.0019         Pr(|T| > |t|) = 0.0038          Pr(T > t) = 0.9981
```

We see that giving in the give-or-take game is lower on average, and the average difference between treatments is −0.62. Moreover, there is strong evidence that giving is lower in the give-or-take treatment, since the one-tailed(<) p-value is 0.0019.

The non-parametric test appropriate in this situation is the Wilcoxon signed ranks test (see Siegel & Castellan, 1988). As with the parametric test, this test is based on the differences in giving between treatments for each observation. 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 the take-treatment has no effect, these two rank sums should be roughly equal. The test is therefore based on a comparison of these two numbers. The test is performed using the `signrank` command in STATA, as below.

```
. signrank y2=y1

Wilcoxon signed-rank test

      sign |      obs   sum ranks    expected
-----------+---------------------------------
  positive |        3       116.5         340
  negative |       13       563.5         340
      zero |       34         595         595
-----------+---------------------------------
       all |       50        1275        1275

unadjusted variance     10731.25
adjustment for ties        -7.50
adjustment for zeros   -3421.25
                       ----------
adjusted variance        7302.50

Ho: y2 = y1
           z =  -2.615
    Prob > |z| =   0.0089
```

The rank sum for the negative differences is clearly a higher number, at 563.5. The test gives a (two-tailed) p-value of 0.0089 which represents strong evidence that giving is different under the take treatment. The one-tailed p-value is 0.0045. This p-value is, as expected, larger than the 0.0019 obtained above from the corresponding parametric test, indicating that the evidence from this non-parametric test is less strong.

Actually, the point must be made that 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 can also be peformed in STATA:

```
. signtest y2=y1

Sign test

        sign |    observed    expected
-------------+------------------------
    positive |           3           8
    negative |          13           8
        zero |          34          34
-------------+------------------------
         all |          50          50

One-sided tests:
  Ho: median of y2 - y1 = 0 vs.
  Ha: median of y2 - y1 > 0
      Pr(#positive >= 3) =
         Binomial(n = 16, x >= 3, p = 0.5) =  0.9979

  Ho: median of y2 - y1 = 0 vs.
  Ha: median of y2 - y1 < 0
      Pr(#negative >= 13) =
         Binomial(n = 16, x >= 13, p = 0.5) =  0.0106

Two-sided test:
  Ho: median of y2 - y1 = 0 vs.
  Ha: median of y2 - y1 != 0
      Pr(#positive >= 13 or #negative >= 13) =
         min(1, 2*Binomial(n = 16, x >= 13, p = 0.5)) =  0.0213
```

The relevant p-value is the second one, 0.0106. Again there is evidence that giving is lower under the take treatment. However, being the result of the most non-parametric of tests, the evidence is weaker still, and since the p-value is greater than 0.01, the evidence can no longer be classified as strong.

In the discussion of within-subject designs in Chapter 2, the problem of order effects was raised. The problem may be relevant to the current situation. It might be expected that a subject's experience of the give-only game somehow influences their behaviour in the subsequent give-or-take game, and hence the treatment effect is confounded by the order of the two treatments. As mentioned in Chapter 2, a way of addressing such a concern is to use a crossover design; that is, a design in which half of the subjects are presented with the same two treatments but in the reverse order. Any order effect could then be controlled for in the process of treatment testing. This is in fact the approach taken by Chlaß & Moffatt (2012).

## 3.8 Summary and Further Reading

This chapter has attempted to cover a wide variety of treatment tests, with examples. Readers wishing to learn more about any of these tests are directed to Siegel & Castellan (1988). Camerer (2003, ch. 2) surveys a large number of treatment tests applied to behaviour in ultimatum games, dictator games and trust games.

An important decision is the choice between non-parametric and parametric approaches, and this decision is often guided by the scale of measurement of the

data (nominal, ordinal, or cardinal). There is a large literature on this point, including Harwell & Gatti (2001).

One particular application of treatment testing that has been covered is the testing of gender effects. Croson & Gneezy (2009) provide a thorough review of the literature on the testing of gender effects in economic experiments.

The within-subject testing approach was applied to a number of situations including the preference reversal (PR) phenomenon. The PR phenomenon was first introduced to the economics literature by Grether & Plott (1979). A very useful survey of research on the phenomenon was later provided by Seidl (2002).

Another technique covered in this chapter is the bootstrap. Readers seeking more detail on this technique are referred to Efron & Tibshirani (1993) and MacKinnon (2002).

## Exercises

1. Burnham (2003) considers the binary decision to give nothing using the binomial test. Reproduce the results appearing in his Table 4.
2. Eckel & Grossman (1998) present a large number of tests of the effect of gender on dictator game giving. Their data set is presented within the article. Reproduce as many of their results as you can.
3. Branas-Garza (2007) investigate framing effects in dictator game giving. Dictators in the treatment group (2) had an additional line at the end of the instructions, which, translated from Spanish into English, means roughly "Note that your recipient relies on you". Dictators in the control group (1) did not see such a line. The experiment was performed both in a classroom (C) and in a lab (L). The distributions of contributions for each group are shown in the table:

| Donation | C1 | C2 | L1 | L2 |
|---|---|---|---|---|
| 0 | 11 | 2 | 5 | 1 |
| 1 | 1 | 0 | 2 | 2 |
| 2 | 4 | 3 | 6 | 3 |
| 3 | 2 | 7 | 6 | 5 |
| 4 | 1 | 5 | 5 | 10 |
| 5 | 1 | 3 | 3 | 5 |
| Total | 20 | 20 | 27 | 26 |

   (a) Create a data set with 93 rows (1 row for each subject), and create the following three variables:
   Setting: 1 if classroom; 2 if lab
   Treatment: 1 if additional sentence not seen; 1 if seen
   X: contribution

   (b) Reproduce their results.

# Chapter 4
# Theory Testing, Regression, and Dependence

## 4.1 Introduction

This chapter is mainly concerned with the role of experimental data in theory testing. As stressed at the start of Chapter 1, the focus of many economic experiments is on the functioning of a particular economic institution, and the objective is to test a particular economic theory in the context of that institution. The two institutions used as examples in this chapter are auctions and contests. For both of these institutions, the economic theory is very well developed, and leads to very clear "fundamental predictions", often in the form of the "risk-neutral Nash equilibrium prediction". This fundamental prediction provides a natural starting point for the range of tests demonstrated in this chapter.

In the settings considered in this chapter, and indeed in many experimental settings, behaviour of experimental subjects tends to depart in systematic ways from the "fundamental prediction" of theory. In the examples seen later, these departures take the form of systematic "over-bidding" relative to the Nash equilibrium prediction in auctions and contests. Hence, if our only objective were to test the fundamental predictions of Nash equilibrium theories, this objective would be straightforwardly met: the theory would be rejected. However, there are other levels at which theory may be tested. A typical theory gives rise to a number of "comparative static predictions". An example that is prominent in this chapter is the effect of changing the number of bidders in an auction. If the theory predicts that an increase in the number of bidders causes a decrease in bids, this is a comparative static prediction that can easily be tested using experimental data, simply by comparing the levels of bids between treatments with different numbers of bidders.

The principal objective of the chapter is to demonstrate how such comparative static predictions of the theory, and other types of treatment effect, may be tested within the framework of a linear regression model. The basic idea is that a treatment test can be performed as a test of significance of a "treatment dummy" in a linear regression model whose dependent variable is the outcome variable for the test. In fact, it can be shown that when a regression is performed with only a single dummy variable and an intercept, the t-test for the significance of the effect of the dummy

is equivalent to the independent samples t-test, covered in Chapter 3, for testing the difference in means between the treatment and control groups.

The use of regression analysis for the purposes of treatment testing has a number of major advantages. Firstly, it is possible to test the effects of more than one treatment simultaneously, and, if necessary, interactions between them. Secondly, it is possible to control for the effects of other variables (e.g. subject characteristics) that might affect the outcome. Thirdly, it is possible to adjust for *dependence* between observations. Dependence normally takes the form of clustering, either at the level of the individual subject, or at the level of the group of subjects, or at the level of the experimental session. When a straightforward linear regression is being performed, an adjustment may easily be made to the standard errors to make them "cluster-robust", hence validating the treatment tests. Of course, a superior approach is fully to respect the "panel" structure of the data by using panel estimators, that are fully efficient. If there is more than one level of clustering (e.g. subject and session), a fully efficient estimator is one that follows the multi-level modelling approach. All of these methods are covered in this chapter.

There are several reasons for the choice of experimental auctions as the principal context in which the treatment tests are demonstrated. Firstly, as mentioned above, auction theory is very well developed (see Krishna, 2010) and gives rise to many clear predictions which may be tested econometrically. The central prediction is the risk-neutral Nash equilibrium (RNNE) bid function. Secondly, while it is possible to test these predictions using data on real auctions (see for example Laffont et al., 1995), the econometrics of experimental auctions is much simpler. This is because experimental auctions are conducted using the *induced value* methodology, in which, in any given round of the experiment, each subject is given a "private value" or "signal" by the experimenter, and makes their bidding decision on this basis. In other words, the private values are fully known by the investigator. This is unlike real auction data, in which private values are clearly unobserved, and estimation of their distribution presents a major obstacle to estimation and testing. With experimental data, because the private values are known, adherence or otherwise to the RNNE can be tested directly using a one-sample test, and certain comparative static predictions of auction theory can be tested using two-sample treatment tests. Also, the "bid function", that is, the equation showing bid as a function of private value, can be estimated using linear regression analysis, since both the dependent and explanatory variables are fully observed. This regression can of course be made to include treatment dummies so that the comparative static predictions of interest may be tested. Thirdly, the regression framework also allows the investigation of other determinants of the bid, such as experience and accumulated balance. Fourthly, experimental auction data has a very clear panel structure, with clustering at the level of bidder, and at the level of the experimental session, and so provides the ideal setting in which to demonstrate the panel data techniques described in the last paragraph.

When data from experimental auctions are analysed, there is an almost universal tendency for bidding to be more "aggressive" (i.e. higher) than predicted by RNNE. There are a number of explanations for this phenomenon, the dominant one being failure to adjust for the fact that the highest private signal is likely to be

too high, this failure being known as the "winner's curse". Other explanations for over-bidding include: joy of winning the auction; use of simple heuristics; regret; confusion; experimenter demand effect; and house money effect. Some of these explanations are considered in applications of the testing techniques developed later.

The other context used for illustration is that of the contest experiment. Contest experiments are similar to auction experiments in some ways: it is straighforward to compute the RNNE, and there is a tendency to over-bid relative to the RNNE. However, the treatment tests we apply to contest data have a different sort of objective. With auctions, we are mainly interested in testing comparative static predictions *of the theory*. With contests, the emphasis shifts to an investigation of the "drivers of out-of-equilibrium play" – that is, which features *of the experimental design* lead to behaviour that is closer to, or further away from, the RNNE.

We also use the context of contest experiments to demonstrate the technique of meta-analysis, which provides yet another means of testing the predictions of theory and of identifying the drivers of out-of-equilibrium play.

Section 4.2 provides a minimal overview of auction theory, intended simply to introduce the concepts required for an understanding of the various theoretical predictions that are tested later in the chapter. Simulated experimental auction data is used in all of the sections that follow. Section 4.3 considers some basic tests of the fundamental predictions of auction theory. Section 4.4 considers tests of camparative static predictions, both as standard treatment tests and as tests within regression models. It is also explained how dependence can be accommodated. Section 4.5 extends the model to a multiple regression context, allowing for the effects of the level of uncertainty, the accumulation of bidding experience, and the accumulation of cash balances. Section 4.6 introduces panel data estimators and applies them to the models of Section 4.5. Section 4.7 demonstrates how multilevel modelling can be used to accommodate both subject-level and session-level dependence. Section 4.8 introduces contest experiments, and demonstrates some tests in this context. Section 4.9 considers a meta-analysis of a sample of published results from contest research. Section 4.10 summarises the chapter.

## 4.2 Experimental Auctions

### *4.2.1 Overview of auction theory*

There are two types of auction: common-value auctions and private-value auctions. In a common-value auction, each bidder submits a bid for a particular object which has the same value to all bidders, but this value is unknown. Since the value is unknown, a bidder can make a loss if they win the auction and then find that the price they are paying is higher than the true value of the object.

In a private-value auction, each bidder has a different value for the object, and this value is known to the bidder. All the bidder needs to do is make a bid lower than their own private value. That way, they make a profit, if they win the auction.

However, bidding too far below the private value clearly reduces the probability of winning.

The winner of the auction is the one who submits the highest bid. But what price does this bidder pay? It might seem obvious that they should pay the price that they bid. If they do, they are playing according to the rules of a "first-price" auction. However, some auctions, labelled "second price" auctions, are designed so that the winning bidder pays whatever price was bid by the *second highest* bidder. One important reason why we are interested in second-price auctions is that they are strategically equivalent to English (increasing) auctions. This is the type of auction with which we are most familiar, with an auctioneer starting with a low price and gradually increasing the price until only one bidder remains. Obviously the highest bidder will stop bidding immediately after the second highest bidder drops out.

A first-price (sealed-bid) auction is strategically equivalent to the less familiar "Dutch auction": an auctioneer starts with a high price and gradually lowers the price until it is accepted by the highest bidder.

One of the most remarkable results in auction theory is the *revenue-equivalence theorem*: with risk-neutral bidders, the expected price paid under both auctions (first-price and second-price) is the same.

Here, we are interested in the "bid function". That is the function $b(x)$ that tells us what the bid should be if the private signal (or private value) is x. Given any auction type, auction theory may be applied to predict the bid function. The bid function derived from the theory is usually the RNNE bid function.

One RNNE bid function in which we are particularly interested is that arising in a second-price common-value auction. As shown by Kagel et al. (1995), this bid function is:

$$b(x) = x - \frac{\epsilon(N-2)}{N} \tag{4.1}$$

where $N$ is the number of bidders, and $\epsilon$ is a measure of the level of uncertainty implicit in the private signals. The latter will be explained fully in due course.

Naturally, bids are predicted to be lower than the private signal. According to (4.1), the amount by which they are lower than the private signal clearly depends positively on the level of uncertainty ($\epsilon$) and positively on the number of bidders ($N$). These are comparative static predictions that will be tested in later sections by applying a variety of different testing techniques to the (simulated) using of data on bids.

The amount by which the bid is lower than the private signal is known as the "bid factor". We will label the bid factor "y". In the context of a second-price common-value auction, the (RNNE) bid factor is, from (4.1):

$$y = x - b(x) = \frac{\epsilon(N-2)}{N} \tag{4.2}$$

A very important feature of bidding behaviour in common-value auctions is "winner's curse". This is the tendency for subjects to fail fully to take into consideration the adverse selection problem: winning the lottery often implies that the private signal was "too high". Failure to adjust fully for this means that observed

bidding is usually somewhat higher than the RNNE prediction (4.1), and hence that the bid factor is somewhat *lower* than the RNNE bid factor (4.2).

### 4.2.2 Carrying out an experimental auction

We will describe the experimental methodology relevant to a second-price, common-value auction.

Subjects are recruited to sessions consisting of a series of auction periods. Because "stranger matching" is desirable, the number of subjects in a session must be greater than the number required for a single bidding group, and, in each round, different bidding groups are formed as random combinations of subjects within the session. For a given bidder-group in a given period, the experimenter generates a "true value" for the imaginary object, $x_0$; this value is not revealed to the bidders. The experimenter then provides each bidder with a "private signal", $x$, drawn from a uniform distribution on $[x_0 - \epsilon, x_0 + \epsilon]$. Bidders know the value of $\epsilon$, which is a measure of bidder uncertainty. But while they know their own private signal, they do not know the private signals of other bidders. Each subject in the bidding group submits a sealed bid for the item.

Each subject is given an initial balance at the start of the experiment. For a given bidder-group in a given period, the winner of the auction is the bidder with the highest bid. The winning bidder buys the object at a price equal to the bid of the *second highest* bidder. At this point, the true value of the object is revealed. The winning bidder receives a profit which is the true value minus the price paid. It is possible that this profit is negative. Other bidders receive a profit of zero for that round. In each round, each bidder's profit is added to their existing balance. Any subject whose balance falls below zero is declared "bankrupt" and is excluded from further rounds. Bankrupt subjects only receive a "participation fee" at the end of the session. Subjects who survive to the end of the session are paid their end-of-experiment balance as well as their participation fee.

### 4.2.3 The simulated auction data

Data have been simulated from a second-price common-value auction.[1] The simulated data is contained in the file **common_value_sim**. There are a total of 160 subjects, divided into 16 sessions. Each subject experiences 30 auction periods. The true values ($x_0$) have been drawn from a uniform distribution on [25, 975].

[1] The design assumed in the simulation bears strong similarities to the designs of Kagel et al. (1995) and Ham et al. (2005), although both of those are private value auctions – here we consider common value auctions.

Then, with each drawn $x_0$, a set of "private signals", $x$, have been drawn from a uniform distribution on $[x_0 - \epsilon, x_0 + \epsilon]$, with the parameter $\epsilon$ varying between sessions. Between the 16 sessions, values of N and $\epsilon$ vary, as follows:

| Sessions | # subjects | $N$ | $\epsilon$ | RNNE bid function | RNNE bid factor | Observations |
|---|---|---|---|---|---|---|
| 1–4 | 8 | 4 | 12 | $b(x) = x - 6$ | 6 | 871 |
| 5–8 | 12 | 6 | 12 | $b(x) = x - 8$ | 8 | 1250 |
| 9–12 | 8 | 4 | 24 | $b(x) = x - 12$ | 12 | 840 |
| 13–16 | 12 | 6 | 24 | $b(x) = x - 16$ | 16 | 1273 |

For each of Sessions 1–4 and 9–12, there are eight subjects, and in each round, these subjects are divided into two groups of $N = 4$ bidders. In each of Sessions 5–8 and 13–16, 12 subjects are divided into two groups of $N = 6$ bidders.

Also given in the table are the RNNE bid functions and RNNE bid factors, for each group of sessions. Note that, in accordance with (4.1), these are different for each group of sessions, by virtue of the values of $N$ and $\epsilon$ changing between groups of sessions.

The design is a $2 \times 2$ *full-factorial* design, since it includes all four of the possible combinations of $N$ and $\epsilon$. This is essential for a proper test of the theory, since the theoretical bid function contains an interaction effect between $N$ and $\epsilon$, in addition to the two main effects.

In the simulation, each subject starts with a balance of 14 experimental units. The bids are simulated in such a way as to depend in expected ways on the private signal, the period number, the number of players, the level of uncertainty, and the cumulative balance of the subject. The total numbers of available observations (shown in the final column of the table) varies between treatments, for the following reasons. Treatments with $N = 6$ yield more data than those with $N = 4$. Also, when a subject goes bankrupt, they are excluded from further rounds so no further observations are generated from them. Finally, observations for which the private signals are outside of the range [60, 963] are excluded from estimation, because the RNNE bid function is more complicated at the ends of the distribution.

Figure 4.1 shows a screenshot of the first 31 rows of the data. It is useful to focus on one of these rows. The fifth row contains the following information. In the fifth period of the first session, subject 1 participated in market 1 (remember that in each period the group of subjects are divided into two markets), the true value of the object is 592, and subject 1 receives a private signal of 591. Subject 1 bids 587, so that her bid factor is $591 - 587 = 4$. With this bid, the subject wins the auction (indicated by winner $= 1$), and pays the second price (spr) which is 584. She earns a profit of $592 - 584 = 8$, which is added to her current balance of 14, making her balance 22 in the following period. Note from other rows that this subject wins the auction in 5 of the 30 rounds, and accumulates a balance of 38 at the end of the 30 rounds.

In order to examine particular auctions, it is more convenient to sort the data using the command `sort session period market i`. The sorted data set is shown in Figure 4.2. To see what happened in period 1 of session 1, we examine the first four rows. We see that subjects 2, 3, 4, and 6 were selected to take part in

have been drawn from a uni-
$\epsilon$ varying between sessions.
lows:

| bid factor | Observations |
|---|---|
| 6 | 871 |
| 8 | 1250 |
| 12 | 840 |
| 16 | 1273 |

ects, and in each round, these
In each of Sessions 5–8 and
6 bidders.
s and RNNE bid factors, for
(4.1), these are different for
l $\epsilon$ changing between groups

includes all four of the pos-
a proper test of the theory,
n effect between $N$ and $\epsilon$, in

ce of 14 experimental units.
n expected ways on the pri-
the level of uncertainty, and
ers of available observations
treatments, for the following
an those with $N = 4$. Also,
further rounds so no further
vations for which the private
from estimation, because the
the distribution.
s of the data. It is useful to
following information. In the
n market 1 (remember that in
o markets), the true value of
al of 591. Subject 1 bids 587,
the subject wins the auction
(spr) which is 584. She earns
nt balance of 14, making her
ows that this subject wins the
lance of 38 at the end of the

e convenient to sort the data
`t i`. The sorted data set is
l 1 of session 1, we examine
were selected to take part in

| | session | i | period | market | x_0 | x | bid | winner | spr | profit | balance |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 596 | 589 | 582 | 0 | 602 | 0 | 14 |
| 2 | 1 | 1 | 2 | 2 | 375 | 378 | 370 | 0 | 370 | 0 | 14 |
| 3 | 1 | 1 | 3 | 2 | 556 | 545 | 548 | 0 | 551 | 0 | 14 |
| 4 | 1 | 1 | 4 | 2 | 945 | 950 | 942 | 0 | 942 | 0 | 14 |
| 5 | 1 | 1 | 5 | 1 | 592 | 591 | 587 | 1 | 584 | 8 | 14 |
| 6 | 1 | 1 | 6 | 1 | 387 | 395 | 391 | 1 | 386 | 1 | 22 |
| 7 | 1 | 1 | 7 | 1 | 754 | 744 | 743 | 0 | 752 | 0 | 23 |
| 8 | 1 | 1 | 8 | 1 | 744 | 735 | 734 | 0 | 750 | 0 | 23 |
| 9 | 1 | 2 | 9 | 2 | 661 | 670 | 661 | 0 | 661 | 0 | 23 |
| 10 | 1 | 2 | 10 | 1 | 913 | 903 | 899 | 0 | 905 | 0 | 23 |
| 11 | 1 | 2 | 11 | 2 | 927 | 933 | 926 | 0 | 926 | 0 | 23 |
| 12 | 1 | 1 | 12 | 1 | 370 | 359 | 352 | 0 | 374 | 0 | 23 |
| 13 | 1 | 1 | 13 | 1 | 727 | 719 | 713 | 0 | 716 | 0 | 23 |
| 14 | 1 | 2 | 14 | 2 | 703 | 712 | 709 | 1 | 708 | -5 | 23 |
| 15 | 1 | 1 | 15 | 1 | 311 | 303 | 297 | 0 | 298 | 0 | 18 |
| 16 | 1 | 1 | 16 | 2 | 139 | 142 | 138 | 0 | 138 | 0 | 18 |
| 17 | 1 | 2 | 17 | 2 | 65 | 67 | 58 | 0 | 58 | 0 | 18 |
| 18 | 1 | 1 | 18 | 1 | 438 | 437 | 435 | 0 | 439 | 0 | 18 |
| 19 | 1 | 1 | 19 | 2 | 203 | 194 | 196 | 0 | 196 | 0 | 18 |
| 20 | 1 | 1 | 20 | 2 | 531 | 533 | 525 | 0 | 529 | 0 | 18 |
| 21 | 1 | 1 | 21 | 1 | 328 | 332 | 324 | 0 | 326 | 0 | 18 |
| 22 | 1 | 2 | 22 | 1 | 575 | 586 | 578 | 1 | 565 | 10 | 18 |
| 23 | 1 | 2 | 23 | 2 | 765 | 756 | 748 | 0 | 751 | 0 | 28 |
| 24 | 1 | 2 | 24 | 2 | 519 | 509 | 508 | 0 | 513 | 0 | 28 |
| 25 | 1 | 2 | 25 | 2 | 507 | 500 | 495 | 0 | 505 | 0 | 28 |
| 26 | 1 | 2 | 26 | 2 | 487 | 484 | 482 | 0 | 483 | 0 | 28 |
| 27 | 1 | 1 | 27 | 2 | 709 | 714 | 711 | 0 | 711 | 0 | 28 |
| 28 | 1 | 2 | 28 | 1 | 779 | 787 | 783 | 1 | 769 | 10 | 28 |
| 29 | 1 | 1 | 29 | 2 | 697 | 685 | 676 | 0 | 687 | 0 | 38 |
| 30 | 1 | 1 | 30 | 2 | 185 | 180 | 172 | 0 | 184 | 0 | 38 |
| 31 | 1 | 2 | 1 | 1 | 705 | 717 | 713 | 1 | 698 | 7 | 14 |

Figure 4.1: Common value auction data

| | session | i | period | market | x_0 | x | bid | winner | spr | profit | balance | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 7 | 1 | 1 | 705 | 717 | 713 | 1 | 698 | 7 | 14 | 4 |
| 2 | 1 | 8 | 1 | 1 | 705 | 700 | 696 | 0 | 698 | 0 | 14 | 4 |
| 3 | 1 | 4 | 1 | 1 | 705 | 699 | 697 | 0 | 698 | 0 | 14 | 2 |
| 4 | 1 | 6 | 1 | 1 | 705 | 705 | 698 | 0 | 698 | 0 | 14 | 7 |
| 5 | 1 | 1 | 1 | 2 | 596 | 589 | 582 | 0 | 602 | 0 | 14 | 7 |
| 6 | 1 | 5 | 1 | 2 | 596 | 595 | 582 | 0 | 602 | 0 | 14 | 13 |
| 7 | 1 | 7 | 1 | 2 | 596 | 605 | 602 | 0 | 602 | 0 | 14 | 3 |
| 8 | 1 | 8 | 1 | 2 | 596 | 605 | 603 | 1 | 602 | -6 | 14 | 2 |
| 9 | 1 | 1 | 2 | 1 | 303 | 302 | 299 | 0 | 309 | 0 | 21 | 12 |
| 10 | 1 | 4 | 2 | 1 | 303 | 299 | 293 | 0 | 309 | 0 | 14 | 1 |
| 11 | 1 | 7 | 2 | 1 | 303 | 313 | 309 | 0 | 309 | 0 | 14 | 4 |
| 12 | 1 | 8 | 2 | 1 | 303 | 315 | 312 | 1 | 309 | -6 | 8 | 3 |
| 13 | 1 | 1 | 2 | 2 | 375 | 378 | 370 | 0 | 370 | 0 | 14 | 8 |
| 14 | 1 | 5 | 2 | 2 | 375 | 372 | 368 | 0 | 370 | 0 | 14 | 4 |
| 15 | 1 | 4 | 2 | 2 | 375 | 373 | 372 | 1 | 370 | 5 | 14 | 1 |
| 16 | 1 | 6 | 2 | 2 | 375 | 379 | 362 | 0 | 370 | 0 | 14 | 17 |
| 17 | 1 | 2 | 3 | 1 | 442 | 433 | 428 | 0 | 430 | 0 | 14 | 5 |
| 18 | 1 | 6 | 3 | 1 | 442 | 448 | 441 | 1 | 430 | 12 | 14 | 7 |
| 19 | 1 | 7 | 3 | 1 | 442 | 432 | 430 | 0 | 430 | 0 | 14 | 2 |
| 20 | 1 | 8 | 3 | 1 | 442 | 446 | 441 | 1 | 430 | 12 | 2 | 5 |
| 21 | 1 | 1 | 3 | 2 | 556 | 545 | 548 | 0 | 551 | 0 | 14 | -3 |
| 22 | 1 | 2 | 3 | 2 | 556 | 559 | 551 | 0 | 551 | 0 | 21 | 8 |
| 23 | 1 | 4 | 3 | 2 | 556 | 563 | 560 | 1 | 551 | 5 | 19 | 3 |
| 24 | 1 | 5 | 3 | 2 | 556 | 560 | 547 | 0 | 551 | 0 | 14 | 13 |
| 25 | 1 | 2 | 4 | 1 | 322 | 333 | 317 | 0 | 320 | 0 | 21 | 16 |
| 26 | 1 | 4 | 4 | 1 | 322 | 321 | 320 | 0 | 320 | 0 | 24 | 1 |
| 27 | 1 | 7 | 4 | 1 | 322 | 331 | 326 | 1 | 320 | 2 | 14 | 5 |
| 28 | 1 | 8 | 4 | 1 | 322 | 316 | 313 | 0 | 320 | 0 | 14 | 3 |
| 29 | 1 | 1 | 4 | 2 | 945 | 950 | 942 | 0 | 942 | 0 | 14 | 8 |
| 30 | 1 | 3 | 4 | 2 | 945 | 945 | 940 | 0 | 942 | 0 | 14 | 5 |
| 31 | 1 | 5 | 4 | 2 | 945 | 938 | 924 | 0 | 942 | 0 | 14 | 14 |

Figure 4.2: Common value auction data: re-sorted using "sort session period market i"

"market 1". The true value was 705, and the auction was won by subject 2 with a bid of 713. This subject pays a price 698, the bid of subject 6, this being the second highest bid. The profit earned by subject 2 is $705 - 698 = 7$.

## 4.3 Tests of Auction Theory

### *4.3.1 A test of RNNE in a second-price common value auction*

In this section we consider direct tests of the fundamental predictions of the theory. These tests are one-sample tests applied to the bid factor. An important point is that when conducting these tests, we are making the assumption of independence of observations, an assumption which, as we will see many times later, is hard to justify. Hence the results obtained in this section might not be taken too seriously. On several occasions later in the chapter, the fundamental prediction will be again tested, but within the context of regression. There, appropriate adjustments are made for dependence in the data simply because it is straightforward to make such adjustments in the regression framework. Those regression-based tests will provide more reliable conclusions regarding the data's closeness to theoretical predictions.

Here, we consider the two benchmark models used by Kagel et al. (1995): a naïve bidding model, and the RNNE. The first model represents an extreme form of naïvety. The second represents an extreme form of rationality. It is not anticipated that observed behaviour will correspond closely to either of the two models.

Let us focus on Sessions 1–4, in which the number of bidders ($N$) is fixed at 4, and the uncertainty parameter ($\epsilon$) is fixed at 12. If we insert these values into (4.1) we observe that the RNNE prediction in these sessions is simply:

$$b(x) = x - 6 \tag{4.3}$$

Another way of stating this prediction is that the bid factor ($y$) equals 6. In Figure 4.3, we show a histogram of the bid factor for sessions 1–4, with a vertical line at the RNNE prediction of 6. The histogram shows that there is a good deal of variation in the bid factor, but it is fairly clear that the main part of the distribution is to the left of the RNNE prediction of 6.

An obvious way to conduct a formal statistical test of the theory is to test the null hypothesis that the population mean of $y$, $\mu$ say, is equal to 6.0, against the alternative hypothesis that it equals some value less than 6.0 (since values less than 6.0 arise as a result of "winner's curse"). If there are $n$ observations in the data set we would first find the mean and standard deviation of the bid factor, $\bar{y}$ and $s$, and we would compute the one-sample t-test statistic:

$$t = \frac{\bar{y} - 6.0}{s/\sqrt{n}} \tag{4.4}$$

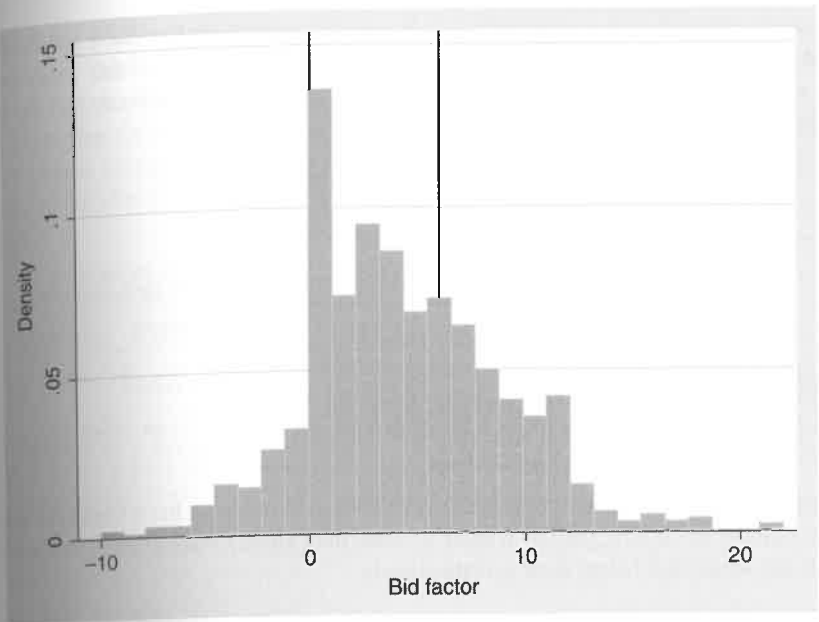



Figure 4.3: A histogram of the bid factor for sessions 1–4. Vertical lines drawn at RNNE prediction (6) and naïve prediction (0)

It is easier to do this test in STATA:

```
. * USE ONLY FIRST 4
SESSIONS (TO FIX N=4 AND EPSILON=6)

. keep if session<5
(1349 observations deleted)


. * test of RNNE

. ttest y=6

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       y |     871    4.337543      .1535      4.5302    4.036269    4.638817
------------------------------------------------------------------------------
    mean = mean(y)                                                t = -10.8303
Ho: mean = 6                                    degrees of freedom =      870

    Ha: mean < 6                 Ha: mean != 6                 Ha: mean > 6
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

We see that the mean bid factor is 4.34, but this is significantly below 6.0, as is clear from the p-value of 0.0000. We therefore strongly reject the RNNE prediction, and we conclude that "winner's curse" is taking hold (since bids are, on average, higher than predicted by RNNE).

The naive bidding model is one in which bidders simply bid their own private signal, implying that the bid factor is zero. This is clearly a very extreme form of

winner's curse. Another vertical line in Figure 4.3 represents this naïve prediction, and it is clear that the majority of the distribution is to the right of this prediction. To conduct a formal test of this naïve model we test the null hypothesis that the bid factor has a mean of zero:

```
. * test of naive behaviour

. ttest y=0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       y |     871    4.337543      .1535      4.5302    4.036269    4.638817
------------------------------------------------------------------------------
    mean = mean(y)                                t =  28.2576
Ho: mean = 0                     degrees of freedom =      870

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

This model is also strongly rejected (p=0.0000). Although we have found evidence that bidders are falling prey to winner's curse, they clearly understand that sensible bids are somewhat below their private signals.

## 4.4 Tests of Comparative Static Predictions

Two very simple models were tested in the last section: the RNNE and the naïve bidding model. Notwithstanding the failure to take account of dependence in the data, both models were decisively rejected. The truth is clearly somewhere in between these two extremes. However, there are many other predictions of the theory that can be tested.

In this section, we consider tests of one of the comparative static predictions of the RNNE theory. Consider the effect on bidding of the number of bidders, $N$. Do we expect bidding behaviour to change if $N$ is changed and all other features of the auction remain the same? Once again appealing to (4.1), we see that the answer is yes. Under RNNE theory, we expect an increase in $N$ to have a negative effect on bids, and therefore a positive effect on the bid factor.

First, we conduct these tests with the assumption of independence between observations. In Section 4.4.3, we shall start to consider methods for allowing for dependence.

### *4.4.1 Standard treatment tests*

The comparative static prediction relating to the number of bidders ($N$) may be tested using standard treatment tests. $N$ takes two different values in the experiment: 4 and 6. The data contains a dummy variable named "N6" taking the value 1 when

$N = 6$ and 0 when $N = 4$. This dummy variable is used as the separation variable for the test. Think of N6 as a "treatment dummy": $N6 = 1$ for the "treatment group" and $N6 = 0$ for the "control group".

Note that for these tests we are using only Sessions 1–8. This is because the other design parameter ($\epsilon$) is fixed over these sessions, as is required if we are to focus on the effect of changes in $N$.

```
. keep if session<9
(2247 observations deleted)

. ranksum y, by(N6)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        N6 |      obs    rank sum    expected
-----------+---------------------------------
         0 |      871    843078.5      924131
         1 |     1250   1407302.5     1326250
-----------+---------------------------------
  combined |     2121     2250381     2250381

unadjusted variance   1.925e+08
adjustment for ties  -984989.31
                     ----------
adjusted variance     1.915e+08

Ho: y(N6==0) = y(N6==1)
           z =  -5.856
    Prob > |z| =   0.0000

. ttest y, by(N6)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     871    4.337543      .1535      4.5302    4.036269    4.638817
       1 |    1250       5.348   .1163788    4.114613     5.11968     5.57632
---------+--------------------------------------------------------------------
combined |    2121     4.93305   .0937551    4.317827    4.749189    5.116912
---------+--------------------------------------------------------------------
    diff |           -1.010457   .1893543               -1.381797   -.6391172
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.3363
Ho: diff = 0                                     degrees of freedom =     2119

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Both the Mann-Whitney test (`ranksum` in STATA) and the independent samples t-test (`ttest` in STATA) result in very strong rejections of the null hypothesis that $N$ has no effect on bids. Also, we must not forget to check the direction of the effect. The second table tells us that when $N = 4$, the mean bid factor is 4.34, while when $N = 6$, the mean bid factor is higher at 5.35. This implies that bids are lower when $N$ increases, and this is what is predicted by the theory. We may therefore conclude that the results of these treatment tests are consistent with the predictions of RNNE theory.

$N = 6$ and 0 when $N = 4$. This dummy variable is used as the separation variable for the test. Think of N6 as a "treatment dummy": $N6 = 1$ for the "treatment group" and $N6 = 0$ for the "control group".

Note that for these tests we are using only Sessions 1–8. This is because the other design parameter ($\epsilon$) is fixed over these sessions, as is required if we are to focus on the effect of changes in $N$.

```
. keep if session<9
(2247 observations deleted)

. ranksum y, by(N6)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        N6 |      obs    rank sum    expected
-----------+---------------------------------
         0 |      871    843078.5      924131
         1 |     1250   1407302.5     1326250
-----------+---------------------------------
  combined |     2121     2250381     2250381

unadjusted variance   1.925e+08
adjustment for ties  -984989.31
                     ----------
adjusted variance     1.915e+08

Ho: y(N6==0) = y(N6==1)
             z =  -5.856
    Prob > |z| =   0.0000

. ttest y, by(N6)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     871    4.337543      .1535     4.5302    4.036269    4.638817
       1 |    1250       5.348   .1163788   4.114613     5.11968     5.57632
---------+--------------------------------------------------------------------
combined |    2121     4.93305   .0937551   4.317827    4.749189    5.116912
---------+--------------------------------------------------------------------
    diff |           -1.010457   .1893543               -1.381797   -.6391172
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.3363
Ho: diff = 0                     degrees of freedom =     2119

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Both the Mann-Whitney test (`ranksum` in STATA) and the independent samples t-test (`ttest` in STATA) result in very strong rejections of the null hypothesis that $N$ has no effect on bids. Also, we must not forget to check the direction of the effect. The second table tells us that when $N = 4$, the mean bid factor is 4.34, while when $N = 6$, the mean bid factor is higher at 5.35. This implies that bids are lower when $N$ increases, and this is what is predicted by the theory. We may therefore conclude that the results of these treatment tests are consistent with the predictions of RNNE theory.

### 4.4.2 Treatment testing using a regression

Treatment tests may also be performed using a regression. In fact, regression on only a dummy variable is equivalent to an independent samples t-test. To verify this, we perform a regression of the bid factor on the dummy N6.

```
. regress y N6

      Source |       SS       df       MS              Number of obs =    2121
-------------+------------------------------           F(  1,  2119) =   28.49
       Model |  524.110821     1  524.110821            Prob > F      =  0.0000
    Residual |  39000.3823  2119  18.4050884            R-squared     =  0.0133
-------------+------------------------------           Adj R-squared =  0.0128
       Total |  39524.4932  2120  18.6436289           Root MSE      =  4.2901

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          N6 |   1.010457   .1893543     5.34   0.000     .6391172    1.381797
       _cons |   4.337543    .145365    29.84   0.000      4.05247    4.622616
------------------------------------------------------------------------------
```

Note that the intercept is 4.34 (the mean of the bid factor in the "control" group) and the coefficient of N6 is 1.01 (the difference in means between the "treatment" and "control" groups, i.e. the effect size). Note also that the t-statistic is identical (in magnitude) to the t-statistic obtained using the independent samples t-test in the last sub-section. Hence the conclusion from this regression test is exactly the same: when N increases, the bid factor increases, meaning that bids decrease.

### 4.4.3 Accounting for dependence: the ultra-conservative test

The tests and regressions performed so far all assume independence between observations. In the current setting (and in most settings in experimental economics) the independence assumption is not met. First of all, since subjects are being observed in a sequence of 30 auction periods, there is dependence within subjects. Since some subjects are more "aggressive" bidders than others (or just more susceptible to winner's curse), their complete set of bids will tend to be higher than those of others. We sometimes describe this phenomenon as "clustering" at the level of the individual subject. There may also be clustering at the level of the session: some sessions may be characterised by aggressive bidding, others by reserved bidding. The presence of clustering means that the tests performed above are invalid.

The most obvious way of dealing with the dependence problem is to start by taking the averages over all dependent observations within each cluster, and then performing standard treatment tests on these averages. The resulting data set has only one observation for each independent unit, and is therefore free of the problem of dependence. The obvious drawback from using this procedure is that the sample size becomes very small, and hence the power of the test is limited. This is the reason why the procedure is described as "ultra conservative"; whatever the observed significance (p-value) of this test, we expect any other test to result in stronger significance (i.e. a smaller p-value).

n. In fact, regression on only
ples t-test. To verify this, we
6.

```
Number of obs =    2121
F(  1,  2119) =   28.48
Prob > F      =  0.0000
R-squared     =  0.0133
Adj R-squared =  0.0128
Root MSE      =  4.2901

------------------------------
     [95% Conf. Interval]
------------------------------
      .6391172    1.381797
      4.05247    4.622616
------------------------------
```

actor in the "control" group
eans between the "treatment"
that the t-statistic is identical
ependent samples t-test in the
ssion test is exactly the same
nat bids decrease.

## -conservative test

independence between obser
experimental economics) the
ce subjects are being observed
ce within subjects. Since some
r just more susceptible to win
higher than those of others. We
" at the level of the individual
he session: some sessions may
erved bidding. The presence of
nvalid.
endence problem is to start by
within each cluster, and then
ges. The resulting data set has
is therefore free of the problem
this procedure is that the sam-
of the test is limited. This is
ra conservative"; whatever the
pect any other test to result in



In the present context, application of the ultra-conservative test involves finding the average bid factor for each session (using the STATA command `table session, contents(n y mean y)`). The resulting means are shown in the following table.

| Sessions | $N$ | $\epsilon$ | RNNE bid factor | Mean bid factors by session |
|---|---|---|---|---|
| 1–4 | 4 | 12 | 6 | 6.25, 5.69, 2.58, 2.36 |
| 5–8 | 6 | 12 | 8 | 7.32, 5.72, 4.31, 3.68 |
| 9–12 | 4 | 24 | 12 | 10.74, 10.73, 9.40, 6.96 |
| 13–16 | 6 | 24 | 16 | 15.78, 12.99, 12.73, 10.63 |

Once again we restrict attention to Sessions 1–8, for which the uncertainty parameter $\epsilon$ is fixed at 12. We have eight independent observations, four with $N = 4$ and four with $N = 6$. We may apply conventional treatment tests to these eight observations, and we obtain the following results.

```
. ranksum mean_y, by(n6)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

          n6 |      obs    rank sum    expected
-------------+---------------------------------
           0 |        4          15          18
           1 |        4          21          18
-------------+---------------------------------
    combined |        8          36          36

unadjusted variance        12.00
adjustment for ties         0.00
                        ----------
adjusted variance          12.00

Ho: mean_y(n6==0) = mean_y(n6==1)
                z =  -0.866
    Prob > |z| =   0.3865

. ttest mean_y, by(n6)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       4        4.22      1.0178      2.0356    .9809063    7.459094
       1 |       4      5.2575    .8090156    1.618031    2.682851    7.832149
---------+--------------------------------------------------------------------
combined |       8     4.73875    .6329902    1.790367    3.241966    6.235534
---------+--------------------------------------------------------------------
    diff |            -1.0375    1.300163               -4.218884    2.143883
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.7980
Ho: diff = 0                                     degrees of freedom =        6

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.2276         Pr(|T| > |t|) = 0.4553          Pr(T > t) = 0.7724
```

Both the Mann-Whitney test and independent samples t-test find no evidence that the number of players has any effect on bids. Remembering that we expect the bid factor to be higher when $N = 6$ than when $N = 4$, we may perform one-tailed tests.

The one-tailed p-values are 0.19 for the Mann-Whitney test (0.38 divided by 2), and 0.23 for the t-test. However, for the reason given above, we must treat these p-values as "ultra conservative", and, on the assumption that a treatment effect is actually present, we anticipate smaller p-values when different types of test are used to test the same hypothesis.

### *4.4.4 Accounting for dependence in a regression*

In Section 4.4.3 we demonstrated a test which explicitly avoids the dependence problem. However, it should be obvious that the procedure is far from satisfactory in view of the large amount of information that is being discarded in the process of averaging observations. Clearly, we wish to utilise a testing procedure that makes use of all of the available information contained in the data set, but which at the same time is robust to the data's dependence structure.

With this objective, we return to the regression framework introduced in Section 4.4.2. In the context of regression, dependence gives rise to a non-diagonal covariance matrix of the error term. This is a violation of one of the classical assumptions of the linear regression model, which requires that the covariance matrix of the error vector is diagonal. The covariance matrix is in fact block-diagonal as a result of the "clustering" of observations by subject or session.

One major advantage of the regression framework for conducting treatment tests is that there are well-established routines for correcting the results to allow for dependence between observations. It is well known (see, for example, Greene, 2008) that whenever the error covariance matrix is non-diagonal, the formula routinely used to compute standard errors is incorrect. There is usually a correct formula, and the choice of formula depends on the nature of the non-diagonality. In this situation of block-diagonality, the appropriate formula is applied by using the `vce(cluster clustvar)` option in STATA, where `clustvar` is the variable specifying to which "cluster" each observation belongs (e.g. the subject identifier, $i$).

Applying this correction to the treatment test peformed in Section 4.4.2, assuming that the clustering is at the level of subject id (i), we obtain the results:

```
. regress y N6, vce(cluster i)

Linear regression                               Number of obs =     2221
                                                F(  1,    79) =     1.90
                                                Prob > F      =   0.1725
                                                R-squared     =   0.0133
                                                Root MSE      =   4.2901

                                    (Std. Err. adjusted for 80 clusters in i)
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          N6 |   1.010457   .7339144     1.38   0.172    -.4503632    2.471277
       _cons |   4.337543   .5999469     7.23   0.000     3.143379    5.531707
------------------------------------------------------------------------------
```

Note that the coefficients are the same as before. All that is changing is the standard errors (and the various quantities that are computed from the standard errors). The standard error of the coefficient of N6 has risen from 0.189 to 0.734 as a result of dealing with clustering. This means that the t-test statistic for the treatment effect is much lower than before, and, unfortunately, the effect is no longer significant (p=0.172).

Adjusting for dependence has transformed a strongly significant treatment effect into an insignificant treatment effect. This example emphatically serves to illustrate the importance of adjusting for dependence in carrying out such tests.

### 4.4.5 *Accounting for dependence: the block bootstrap*

As explained in Chapter 3, "bootstrap tests" are very popular in Experimetrics because they provide a means of carrying out standard parametric tests (e.g. the t-test) without relying on the distributional assumptions that are normally required.

Recall that, in the case of *independent* observations, the bootstrap procedure consists of the following steps:

1. Apply the chosen 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, $\hat{t}^*_j$, $j = 1, \ldots, B$.
3. Compute the standard deviation $s_B$ of the bootstrap test statistics $\hat{t}^*_j$, j=1,. . . , $B$. This will be the bootstrap standard error.

In the presence of clustering, the above procedure will fail, since it fails to replicate the dependence in the data. The *block bootstrap* attempts to replicate the dependence by re-sampling *blocks* of data rather than single observations.

To apply the block bootstrap in STATA, we introduce `bootstrap` into the brackets after the `vce` option. We also specify the number of bootstrap samples. The results are as follows.

```
. reg y N6, vce(bootstrap, rep(999) cluster( i) )
(running regress on estimation sample)

Bootstrap replications (999)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................   50
..................................................  100
..................................................  150
..................................................  200
..................................................  250
..................................................  300
..................................................  350
..................................................  400
..................................................  450
..................................................  500
..................................................  550
..................................................  600
..................................................  650
```

```
.....................................................  700
.....................................................  750
.....................................................  800
.....................................................  850
.....................................................  900
.....................................................  950
.....................................................
.....................................................
```

```
Linear regression                               Number of obs     =       2121
                                                Replications      =        999
                                                Wald chi2(1)      =       1.73
                                                Prob > chi2       =     0.1886
                                                R-squared         =     0.0133
                                                Adj R-squared     =     0.0128
                                                Root MSE          =     4.2901

                                   (Replications based on 80 clusters in i)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          N6 |   1.010457   .7685028     1.31   0.189    -.4957809    2.516695
       _cons |   4.337543   .6176459     7.02   0.000     3.126979    5.548107
------------------------------------------------------------------------------
```

On this occasion, the result from the bootstrap test is not very different from that with the cluster robust standard error in Section 4.4.4.

## 4.5 Multiple Regression with Auction Data

All of the regressions considered so far have contained just one dummy explanatory variable. A further obvious advantage of the regression framework is that many determinants of behaviour can be investigated simultaneously. A model with more than one explanatory variable is a "multiple regression model". To such models we now turn.

### *4.5.1 Introducing the effect of uncertainty*

Recall that the parameter $\epsilon$ is a design parameter which represents the level of uncertainty, in the sense that the higher the value of $\epsilon$, the wider the range of private signals around the true value, and hence the less confidence that an individual bidder can have that his or her private signal is close to the true value. Once again appealing to (4.1), we see that RNNE theory predicts $\epsilon$ to have a negative effect on bids; that is, if the value of $\epsilon$ is increased and nothing else changes, we expect lower bidding. This is another comparative static prediction of RNNE theory. In the experiment there are two different values of $\epsilon$, 12 and 24. This variation enables us to test the comparative static prediction, in a manner similar to our tests of the effect of the number of bidders.

Note that there are now two "treatment variables": N6, which we have already considered, and "eps24", a dummy variable taking the value 1 when $\epsilon = 24$, and

otherwise. At this stage it is useful to reproduce the table shown earlier that summarises the features of the experiment.

| Sessions | Number of subjects | $N$ | $\epsilon$ | RNNE bid function | RNNE bid factor |
|---|---|---|---|---|---|
| 1–4 | 8 | 4 | 12 | $b(x) = x - 6$ | 6 |
| 5–8 | 12 | 6 | 12 | $b(x) = x - 8$ | 8 |
| 9–12 | 8 | 4 | 24 | $b(x) = x - 12$ | 12 |
| 13–16 | 12 | 6 | 24 | $b(x) = x - 16$ | 16 |

With this information, it is possible to write down a ("true") regression model involving both treatment dummies that fully captures the RNNE predictions. This model is:

$$bid = x - 6 - 2.N6 - 6.eps24 - 2.N6 \star eps24 \tag{4.5}$$

For the dependent variable, we are now using the bid itself, not the bid factor. This is possible as long as the private signal ($x$) is included as an explanatory variable, and we obviously expect its coefficient to take the value one. Note that in addition to the two treatment dummies, the equation includes the interaction term $N6 \star eps24$ formed as the product of the two dummies. It will be possible to estimate the coefficient of this interaction variable only because a "full factorial design" has been employed.

Equation (4.5) can be estimated (using the complete data set of all 16 sessions) using a multiple regression. Having estimated the equation, a test command may be used to perform a Wald test of the joint hypothesis that the coefficients are in agreement with the RNNE coefficients given in (4.5). Note that this amounts to a further (more stringent) test of RNNE. This is an F-test, since five separate equalities are being tested simultaneously, one for each of the terms in (4.5).

```
. regress bid x N6 eps24 N6eps24, vce(cluster i)

Linear regression                               Number of obs =      4234
                                                F(  4,   159) =         .
                                                Prob > F      =    0.0000
                                                R-squared     =    0.9997
                                                Root MSE      =    4.3029

                                    (Std. Err. adjusted for 160 clusters in i)
------------------------------------------------------------------------------
             |               Robust
         bid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000129   .0002183  4581.20   0.000     .9996974    1.00056
          N6 |   -1.012588   .7326532    -1.38   0.169    -2.459576    .4343989
       eps24 |   -3.112904   .8103825    -3.84   0.000    -4.713407   -1.512402
    N6eps24 |   -2.661983   1.024279    -2.60   0.010    -4.684929   -.6390368
       _cons |   -4.401871   .6015577    -7.32   0.000    -5.589945   -3.213797
------------------------------------------------------------------------------

. test (x=1) (_cons=-6) (N6=-2) (eps24=-6) (N6eps24=-2)

 ( 1)  x = 1
 ( 2)  _cons = -6
 ( 3)  N6 = -2
 ( 4)  eps24 = -6
 ( 5)  N6eps24 = -2

       F(  5,   159) =   23.01
            Prob > F =    0.0000
```

All of the coefficients appear to have the "correct" signs, and most are significa despite the use of cluster-robust standard errors. In particular, note that when th level of uncertainty ($\epsilon$) rises from 12 to 24, *ceteris paribus*, the bid falls by 5.1 not too far from the prediction of 6.0. In fact, since the 95% confidence interv for this parameter contains the RNNE prediction of 6.0, we may conclude that th comparative static prediction is confirmed by the data.

However, the Wald test that tests all of the RNNE predictions simultaneousl results in a strong rejection of the RNNE hypothesis, with a p-value of 0.0000.

### 4.5.2 *Introducing the effect of experience*

The role of experience is accounted for using the variable representing the perio number. We expect experience to "improve" bidders' performance, in the sense o moving them closer to the RNNE prediction. We also expect a pattern of "conver gence", with bidding reducing steeply in the early periods, but settling down to a stable bidding pattern in later periods. This is captured by including the reciproca of period number, $1/period$, as an explanatory variable. A positive coefficient on this variable will indicate that the learning process takes the form of a trend toward RNNE. The model's other parameters may be interpreted in terms of a long term equilibrium.

Adding $1/period$ (named `rec_period` in the data set) to the regression gives the following results.

```
. regress bid x rec_period N6 eps24 N6eps24, vce(cluster i)

Linear regression                                   Number of obs =     4234
                                                    F(  5,   159) =        .
                                                    Prob > F      =   0.0000
                                                    R-squared     =   0.9997
                                                    Root MSE      =   4.2654

                                    (Std. Err. adjusted for 160 clusters in i)
------------------------------------------------------------------------------
             |               Robust
         bid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000184   .0002142  4670.16   0.000     .9997613    1.000607
  rec_period |    2.911282   .3183578     9.14   0.000     2.282527    3.540036
          N6 |    -1.04156   .7324525    -1.42   0.157     -2.48815    .4050315
       eps24 |   -5.126073   .8112263    -6.32   0.000    -6.728242   -3.523905
    N6eps24 |   -2.619187   1.024981    -2.56   0.012     -4.64352   -.5948539
      _cons |   -4.821608   .6031499    -7.99   0.000    -6.012827   -3.630389
------------------------------------------------------------------------------

. test (x=1) (_cons=-6) (N6=-2) (eps24=-6) (N6eps24=-2)

 ( 1)  x = 1
 ( 2)  _cons = -6
 ( 3)  N6 = -2
 ( 4)  eps24 = -6
 ( 5)  N6eps24 = -2

       F(  5,   159) =   15.81
            Prob > F =    0.0000
```

The variable $1/period$ has a strongly positive coefficient as expected. The test that follows is again a Wald test of RNNE. However, note that this time it is a test of the

closeness of "long run" behaviour (i.e. behaviour after learning) to the RNNE. It is interesting that, although RNNE is again strongly rejected, the rejection is weaker than before, with an F-statistic of 15.8 (compared with 23.0 in the previous model). This is consistent with bidders becoming closer to RNNE with experience.

### 4.5.3 Introducing the effect of cash balance

Recall that subjects start with 14 units, and then profits or losses are earned in each period. "Cash balance" is the amount of money that a subject has accumulated up to the current period. This variable may well have an effect on bidding behaviour for more than one reason. One is "limited liability": since subjects are aware that they will simply be excluded from further rounds if their balance falls below zero, there may be an incentive to bid over-aggressively when the balance becomes low, in a gamble to improve the balance. If this is the case, we expect the cash balance to have a negative effect on bids. However, we might also expect a form of "house money effect": bidders might be expected to bid more aggressively when their balance is high, since this provides a "cushion" against losses, and it is not their money anyway. If this is the case, cash balance is expected to affect bids positively.

When cash balance is included, the results are as follows:

```
. regress bid x balance rec_period N6 eps24 N6eps24, vce(cluster i)

Linear regression                               Number of obs =      4234
                                                F(  6,   159) =         .
                                                Prob > F      =    0.0000
                                                R-squared     =    0.9997
                                                Root MSE      =    4.2602

                                    (Std. Err. adjusted for 160 clusters in i)
------------------------------------------------------------------------------
             |               Robust
         bid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000188   .0002148  4657.23   0.000     .9997643    1.000613
     balance |   -.0124342   .0096541    -1.29   0.200    -.0315011    .0066327
  rec_period |    3.600908   .3841625     6.77   0.000     1.842188    3.359627
          N6 |   -1.158259   .7417146    -1.56   0.120    -2.623142    .3066246
       eps24 |   -4.887118    .842503    -5.80   0.000    -6.551058   -3.223177
     N6eps24 |   -2.730037   1.032137    -2.65   0.009    -4.768504   -.6915695
       _cons |   -4.480833   .6760663    -6.63   0.000    -5.816061   -3.145605
------------------------------------------------------------------------------

. test (x=1) (_cons=-6) (N6=-2) (eps24=-6) (N6eps24=-2)

 ( 1)  x = 1
 ( 2)  _cons = -6
 ( 3)  N6 = -2
 ( 4)  eps24 = -6
 ( 5)  N6eps24 = -2

       F(  5,   159) =    9.57
            Prob > F =    0.0000
```

We see that cash balance has a negative effect, pointing to a "limited liability" effect, but note that this effect is not significant.

## 4.6 Panel Data Estimators

We have been considering methods for dealing with dependence in experimental data. However, so far we have only taken a single step in this direction, by using cluster standard errors. A superior approach is to handle the data set in a panel data framework, which explicitly recognises that $n$ subjects are observed making a decision in each of $T$ time periods. With this approach, it is possible to improve the estimates themselves (i.e. not only the standard errors), using a panel-data estimator instead of OLS. The two most popular panel data estimators are the fixed-effects and random-effects estimators. Both can be represented by the following equation:

$$\begin{aligned} bid_{it} &= \alpha + \beta' x_{it} + \gamma' z_i + u_i + \epsilon_{it} \\ i &= 1 \ldots, n \quad t = 1 \ldots, T \\ Var(u_i) &= \sigma_u^2 \\ Var(\epsilon_{it}) &= \sigma_\epsilon^2 \end{aligned} \tag{4.6}$$

Note that in (4.6) there are two types of explanatory variable. The vector $x_{it}$ contains variables which vary both between subjects and time periods, for example cash balance. The vector $z_i$ contains variables which vary only between subjects, and are fixed over time. Examples of variables appearing in $z_i$ are treatment dummies (where between-subject treatments are applied) and subject characteristics. Note also that there are two error terms: $\epsilon_{it}$ is the conventional equation error term, which is assumed to have mean zero and variance $\sigma_\epsilon^2$ ; $u_i$ is known as the subject-specific term; $u_i$ differs between subjects – hence the $i$-subscript – but it is fixed for a given subject. The two estimators differ in the way this term is interpreted.

The fixed effects estimator is essentially a linear regression which includes a set of $n-1$ dummy variables, one for each subject in the data set (with one excluded to avoid the dummy variable trap). The presence of such dummies has the consequence that the intercept is estimated separately for each subject: the intercept for subject $i$ is $\alpha + u_i$, $i = 1 \ldots, n$.

The random effects estimator does *not* estimate the intercept for each subject. It simply recognises that they are all different, and sets out to estimate only their *variance*, $\sigma_u^2$. Note that random effects is more efficient than fixed effects, because there are far fewer parameters to estimate. We therefore prefer to use random effects if this model turns out to be acceptable.

Another reason for preferring random effects over fixed effects is that the effects of time-invariant variables are not identified under fixed effects. That is, the parameter vector $\gamma$ in (4.6) is not identified under fixed effects. This is important because the variables in which we are most interested, namely the treatment variables, are time-invariant, unless a within-subject design has been used.

Under normal circumstances, to decide between fixed effects (FE) and random effects (RE), the Hausman test is used. In the current situation, however, the FE model is not useful. This is because the variables in whose effects we are most interested (i.e. the treatment dummies N6 and eps24) do not vary within a given subject. Such variation is essential for fixed effects estimation.

The random effects model is, however, useful, and the results are presented below. Panel data commands in STATA can be recognised by the prefix `xt`. For example panel data (linear) regression is carried out using `xtreg`. The fixed effects and random effects estimators are carried out using this command with the options `fe` and `re` respectively. Here we use `re`. Note also that we need to start by declaring the data as panel data using `xtset`, specifying the panel variable followed by the time variable.

```
. xtset i period
       panel variable:  i (unbalanced)
        time variable:  period, 1 to 30, but with gaps
                delta:  1 unit

. xtreg bid x balance rec_period N6 eps24 N6eps24 , re

Random-effects GLS regression                   Number of obs      =      4234
Group variable: i                               Number of groups   =       160

R-sq:  within  = 0.9999                         Obs per group: min =         1
       between = 0.9957                                        avg =      26.5
       overall = 0.9997                                        max =        30

                                                Wald chi2(6)       =  3.22e+07
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         bid |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000459   .0001763  5674.85   0.000      1.000114    1.000805
     balance |   -.0050899    .003371    -1.51   0.131      -.011697    .0015171
  rec_period |    2.453855   .2521596     9.73   0.000      1.959632    2.948079
          N6 |   -1.115177   .7052598    -1.58   0.114     -2.497461    .2671071
       eps24 |   -5.180012   .7765955    -6.67   0.000     -6.702112   -3.657913
    N6eps24 |   -2.594669   .9987159    -2.60   0.009     -4.552116   -.6372215
       _cons |   -4.482867   .5609601    -7.99   0.000     -5.582329   -3.383406
-------------+----------------------------------------------------------------
     sigma_u |  3.0182377
     sigma_e |  2.9870243
         rho |   .50519753   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. test (x=1) (_cons=-6) (N6=-2) (eps24=-6) (N6eps24=-2)

 ( 1)  x = 1
 ( 2)  _cons = -6
 ( 3)  N6 = -2
 ( 4)  eps24 = -6
 ( 5)  N6eps24 = -2

           chi2(  5) =   99.47
         Prob > chi2 =    0.0000
```

Note that the estimates of the two variance components are similar in magnitude: $\hat{\sigma}_u = 3.02$ and $\hat{\sigma}_\epsilon = 2.98$. This implies that between- and within-subject variance are of roughly equal importance. Otherwise, the results are not dissimilar from those of the regression with cluster standard errors presented in Section 4.5.3.

If the between-subject standard deviation, $\sigma_u$, were equal to zero, this would make the random effects model equivalent to a linear regression model. The fact that the estimate of $\sigma_u$ is large in magnitude is strongly suggestive of the superiority of the random effects model. It is possible to make this comparison formally using a likelihood ratio test. This will be done in the context of multi-level modelling in the next section.

## 4.7 Multi-level Modelling

Multi-level modelling is an extension of the random effects framework that allows for more levels of dependence, and also allows for random slopes as well as random intercepts.

The convention adopted here for counting and ordering model levels is similar to that used by Skrondal & Rabe-Hesketh (2004). A "one-level" model is a straightforward linear regression model with a fixed intercept and fixed slopes. For example, imagine that we have $T$ observations, $y_1 \ldots y_T$ on a *single* subject. Then the sample consists of only one cluster, and this is the sense in which there is only one level of clustering. Next, if we have $T$ observations on each of $n$ subjects, $y_{it}, i = 1, \ldots, n, t = 1, \ldots, T$, then a "two-level" model is appropriate, with the subject indicator $i$ representing the second level of clustering. Next, if the $n$ subjects are divided into $J$ sessions, a typical observation is represented by $y_{ijt}$, and a "three-level" model is appropriate, with the session indicator $j$ representing the third (or "highest") level of clustering.

The three-level model just described is specified as follows.

$$\begin{aligned} bid_{ijt} &= \alpha + \beta' x_{it} + \gamma' z_i + u_i + v_j + \epsilon_{ijt} \\ i = 1 \ldots, n \quad j = 1 \ldots, J \quad t = 1 \ldots, T & \\ Var(u_i) &= \sigma_u^2 \\ Var(v_j) &= \sigma_v^2 \\ Var(\epsilon_{it}) &= \sigma_\epsilon^2 \end{aligned} \tag{4.7}$$

In (4.7) $u_i$ is once again the subject-specific random effect, and the new term $v_j$ is the session-specific random effect.

Next, assume that the slope on one of the explanatory variables varies between subjects. For simplicity, let us assume that there is only one variable contained in $x_{it}$, so that $x_{it}$ and the associated parameter $\beta$ are both scalars. The model is:

$$\begin{aligned} bid_{ijt} &= \alpha + \beta x_{it} + \gamma' z_i + u_{0i} + u_{1i} x_{it} + v_j + \epsilon_{ijt} \\ i = 1 \ldots, n \quad j = 1 \ldots, J \quad t = 1 \ldots, T & \\ Var(u_{0i}) &= \sigma_{u0}^2 \\ Var(u_{1i}) &= \sigma_{u1}^2 \\ Var(v_j) &= \sigma_v^2 \\ Var(\epsilon_{it}) &= \sigma_\epsilon^2 \end{aligned} \tag{4.8}$$

In (4.8) there are two between-subject variance parameters: $\sigma_{u0}$ represents between-subject variation in the intercept; $\sigma_{u1}$ represents between-subject variation in the slope on the variable $x_{it}$.

The STATA command for multi-level modelling is `xtmixed`. We will now demonstrate the use of this command in various ways, using the same set of explanatory variables as used in previous sections.

We first consider what would result if the `xtmixed` command contained the list of variables and nothing else. That is:

```
. xtmixed bid x balance rec_period N6 eps24 N6eps24
```

The answer is that this model is the "one-level" model which is identical to the linear regression model, and we would therefore expect coefficient estimates identical to those of the linear regression model as presented in Section 4.5.3.

Next, we introduce clustering at the subject level, giving the "two-level model". Note that this simply requires the addition of "|| $i$ : " at the end of the command line. This model is, of course, equivalent to the random effects model (4.6) and the results below are almost identical to those obtained using the `xtreg` command in Section 4.6. After estimation, the estimates are stored in a vector named "two_level", for later testing.

```
. xtmixed bid x balance rec_period N6 eps24 N6eps24  || i:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood =   -10909.6
Iteration 1:   log likelihood =   -10909.6

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      4234
Group variable: i                               Number of groups   =       160

                                                Obs per group: min =         1
                                                               avg =      26.5
                                                               max =        30

                                                Wald chi2(6)       =  3.24e+07
Log likelihood =   -10909.6                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         bid |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000461   .0001759  5688.01   0.000     1.000116    1.000805
     balance |   -.0050284    .003367    -1.49   0.135    -.0116276    .0015708
  rec_period |    2.450794   .2516203     9.74   0.000     1.957627    2.943961
          N6 |   -1.116107   .7315565    -1.53   0.127    -2.549931    .3177177
       eps24 |   -5.181521   .8053375    -6.43   0.000    -6.759953   -3.603088
    N6eps24 |   -2.593384   1.035923    -2.50   0.012    -4.623756   -.5630121
      _cons |   -4.480923   .5807649    -7.72   0.000    -5.619202   -3.342645
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
i: Identity                  |
                  sd(_cons) |   3.142733   .1845878      2.800995    3.526165
-----------------------------+------------------------------------------------
               sd(Residual) |   2.986586   .0330984      2.922413    3.052167
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  2462.09 Prob >= chibar2 = 0.0000

. est store two_level
```

One piece of output from `xtmixed` that is not present in the `xtreg` output is the "LR test vs linear regression". This is a likelihood ratio (LR) test of the linear regression

model as a restricted version of the random effects model. It is therefore a test for the presence of between-subject variation. The null hypothesis is that the between-subject variation ($\sigma_u^2$) is zero.

The likelihood ratio test is a testing procedure appropriate for nested models. It is computed as:

$$LR = 2(LogL_u - LogL_r) \tag{4.9}$$

where $LogL_u$ and $LogL_r$ are the maximised log-likelihoods from the unrestricted and restricted models respectively. The concept of the log-likelihood function will be explained in detail in Chapter 6. Under the null hypothesis (that the restricted model is true) the test statistic has a $\chi^2(k)$ distribution where $k$ is the number of restrictions. In the present case, only one restriction is being tested ($\sigma_u^2 = 0$), so the null distribution is $\chi^2(1)$. The very high value of the test statistic (2462.09) and the accompanying p-value of 0.0000 emphatically confirm the importance of between-subject variation, and also confirm the superiority of the random effects model over linear regression.

Next, we progress to the "three-level" model by clustering at the session level as well as the subject level. This requires the addition of "|| *session* :" to the command, but note that this must appear *before* "|| *i* :" because it is the higher level of clustering. Placing these two parts of the command the wrong way around will prevent the command from working. We save the results as "three_level".

```
. xtmixed bid x balance rec_period N6 eps24 N6eps24 || session: || i:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -10900.072
Iteration 1:   log likelihood = -10900.072

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      4234

------------------------------------------------------------------------------
                |  No. of       Observations per Group
 Group Variable |  Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------------------
        session |      16        182      264.6        350
              i |     160          1       26.5         30
------------------------------------------------------------------------------

                                                Wald chi2(6)      =  3.24e+07
Log likelihood = -10900.072                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
         bid |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000463   .0001759  5688.39   0.000     1.000118    1.000808
     balance |   -.0042531   .0033643    -1.26   0.206    -.0108471    .0023409
 rec_period |     2.46885   .2515514     9.81   0.000     1.975819    2.961882
          N6 |   -1.108771   1.248898    -0.89   0.375    -3.556566    1.339024
      eps24 |   -5.204657   1.284162    -4.05   0.000    -7.721568   -2.687747
    N6eps24 |   -2.576869   1.766896    -1.46   0.145    -6.039922    .8861837
      _cons |   -4.500616   .9152521    -4.92   0.000    -6.294477   -2.706755
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
  Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
----------------------------+-------------------------------------------------
session: Identity           |
                  sd(_cons) |   1.512163   .3639928       .943422      2.42377
----------------------------+-------------------------------------------------
i: Identity                 |
                  sd(_cons) |   2.756427   .1716992      2.439635     3.114356
----------------------------+-------------------------------------------------
               sd(Residual) |   2.986448   .0330933      2.922285     3.052019
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =  2481.15   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. est store three_level
```

We see that adding the session-level random effect term has caused changes in both the coefficients and the standard errors. Some of the conclusions regarding significance also change. We also see that the estimate of the standard deviation of the session-level random effect ($\sigma_v$) is 1.51, and the confidence interval indicates that this estimate is significantly greater than zero.

Again we can use an LR test to confirm this. This time we are testing the two-level model as a restriction on the three-level model. To carry out this test in STATA, we use the `lrtest` command, as follows.

```
. lrtest three_level two_level

Likelihood-ratio test                                 LR chi2(1)  =     19.06
(Assumption: two_level nested in three_level)         Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on
      the boundary of the parameter space.  If this is not true, then the
      reported test is conservative.
```

As indicated by the note following the test result, this result is "conservative" since the value of ($\sigma_v^2$) under the null hypothesis is zero which is indeed on the boundary of the parameter space. We see that the test results in an overwhelming rejection of the null hypothesis, and we therefore conclude that there is overwhelming evidence of the presence of between-session variation, and that the three-level model should be used in preference to the two-level (random effects) model.

As mentioned at the start of this sub-section, the multi-level modelling approach also allows for random slopes. Let us return to the effect of cash balance, discussed in Section 4.5.3. It was suggested that the effect of cash balance on bids might be negative, for reasons of "limited liability", or it might be positive, due to a form of "house money effect". What we have concluded from all of the models that we have estimated so far is that cash balance has no significant effect on bids. However, it is possible that the effect of cash balance varies over the population, with perhaps some individuals displaying a negative effect, and others displaying a positive effect. In order to investigate this possibility, we allow the slope on cash balance to vary between subjects. Where previously we added "|| $i$ :" at the end of the command to request a random intercept between subjects, we now add

"|| *i* : *balance*" to request a random slope as well as a random intercept. The results are:

```
. xtmixed bid x balance rec_period N6 eps24 N6eps24  || session: || i: balance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -10896.454
Iteration 1:   log likelihood = -10896.046
Iteration 2:   log likelihood = -10896.041
Iteration 3:   log likelihood = -10896.041

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      4234

------------------------------------------------------------------------------
                |  No. of       Observations per Group
 Group Variable |  Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------------------
        session |      16        182      264.6        350
              i |     160          1       26.5         30
------------------------------------------------------------------------------

                                                Wald chi2(6)      =  3.25e+07
Log likelihood = -10896.041                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
         bid |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.000464   .0001756  5697.79   0.000      1.00012    1.000808
     balance |    -.006549   .0043847    -1.49   0.135     -.0151429    .0020449
  rec_period |    2.418536   .2523934     9.58   0.000      1.923854    2.913218
          N6 |   -1.157555   1.243504    -0.93   0.352     -3.594777    1.279668
       eps24 |   -5.239497   1.280784    -4.09   0.000     -7.749787   -2.729207
     N6eps24 |   -2.494244   1.761641    -1.42   0.157     -5.946997    .9585088
       _cons |   -4.415022   .9132111    -4.83   0.000     -6.204883   -2.625161
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
session: Identity            |
                  sd(_cons)  |   1.506631   .3628689      .9397166    2.415555
-----------------------------+------------------------------------------------
i: Independent               |
                sd(balance)  |   .0180286   .0050008      .0104677    .0310504
                  sd(_cons)  |   2.714525   .1734982      2.394912    3.076732
-----------------------------+------------------------------------------------
               sd(Residual)  |   2.975482   .0331918      2.911134    3.041253
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =  2489.21   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. est store random_slope
```

The standard deviation of the random slope, which is represented by $\sigma_{u1}$ in (4.8), is estimated to be 0.018 with a standard error of 0.005. Once again we can use an LR test to test its significance. This time the test compares the three-level model without the random slope ("three_level") against the three-level model with the random slope ("random_slope").

```
. lrtest random_slope three_level

Likelihood-ratio test                                 LR chi2(1)  =      8.06
(Assumption: three_level nested in random_slope)      Prob > chi2 =    0.0045
```

The p-value of the LR test is less than 0.01, suggesting strong evidence that the slope on the cash balance indeed varies between subjects. Since the point estimate of the effect of cash balance is close to zero (−0.0065) we are led to conclude that the population divides fairly evenly between those whose bids depend negatively on cash balance (due to limited liability), and those whose bids depend positively on them (due to a house money effect).

This is a good example of between-subject heterogeneity which is one of the dominant themes of this book. Subjects do respond to changes in their cash balance, but they vary in their responsiveness in such a way that when this heterogeneity is ignored, the effect of cash balance is erroneously estimated to be close to zero.

## 4.8 Modelling Data from Contest Experiments

Partly for the sake of providing a second example of theory testing, in this section we consider the analysis of data from contest experiments.

Contests are similar to auctions in some ways. The essential difference is that in an auction, only the highest bidder can win, while in a contest, any player who bids a positive amount can win. Also, in most auction types (with the exception of the "all-pay auction"), only the winner pays; in contests, all bidders pay.

### *4.8.1 The Tullock contest*

Perhaps the most popular contest model is the "Tullock contest" proposed by Buchanan et al. (1980). It is sometimes referred to as a "rent-seeking contest". In such a contest, there are $n$ players competing for a prize of $v$. Each player $i$ invests an *effort* $e_i$, and player $i$'s probability of winning the prize is defined by the *contest success function* (CSF):

$$p_i(e_i, e_{-i}) = \frac{e_i}{\sum_{j=1}^{n} e_j} \tag{4.10}$$

Given CSF (4.10), the expected payoff for player $i$ is

$$E(\pi_i(e_i, e_{-i})) = p_i(e_i, e_{-i})v - c(e_i) \tag{4.11}$$

where $c(e)$ is the cost of applying effort level $e$. Assuming that the $n$ players are risk neutral and identical to each other, it can easily be shown that the Nash equilibrium effort level for each player ($e^*$) is given by the solution to the following equation:

$$c'(e^*)e^* = \frac{(n-1)}{n^2}v \tag{4.12}$$

In the normal situation in which costs are linear, that is $c(e) = e$, (4.12) simplifies to:

$$e^* = \frac{(n-1)}{n^2}v \tag{4.13}$$

A regular finding in contest experiments is that subjects provide effort levels that are systematically higher than the Nash equilibrium prediction (4.13). That is, as in auction experiments, we observe the phenomenon of "over-bidding".

### *4.8.2 A contest experiment*

A survey of experimental findings on contests has been provided by Sheremeta (2013). Here, we analyse data from a particular contest experiment, conducted by Chowdhury et al. (2014).

The experiment consists of 12 sessions. Each session consists of 12 subjects, each participating in 30 contests. Subjects were matched into groups of $n = 4$, with random rematching after each contest. The value of the prize in all contests was $v = 80$. In each contest, subjects simultaneously selected an effort level between 0 and 80. Applying (4.12) we find that the equilibrium prediction for effort in this situation is $e^* = 15$.

There are two treatments. The first is linear (L) versus convex (C) costs. In sessions employing the linear cost function, the cost of effort was $c(e) = e$ as is standard. In sessions with the convex cost function, the cost of effort was $c(e) = \frac{e^2}{30}$. Because this cost function is such that $c'(15) = 1$, the equilibrium prediction for effort is the same as it is in the linear cost case, that is, 15. The purpose of this treatment is to investigate the extent to which over-bidding is a consequence of the flatness of the payoff function, as suggested (in the context of private value auctions) by Harrison (1989). If this is indeed a reason for over-bidding, we would expect bidding to be lower under the convex costs treatment since the payoff function is steeper under this treatment.

The second treatment is a probabilistic (P) versus share (S) rule for awarding the prize. The probabilistic treatment (P) is the standard situation in which there is a single indivisible prize with winning probabilities given by the CSF (4.10). Under the share treatment (S), the prize is divided between the contestants with shares determined exactly by the CSF (4.10). Again, this treatment does not change the equilibrium prediction of 15. The purpose of this treatment is to investigate the extent to which over-bidding is caused by a non-monetary utility for winning the contest (or "joy of winning"), as suggested by Sheremeta (2010), or by distorted perceptions of probabilities, as suggested by Goeree et al. (2002). If either of these explanations for over-bidding are valid, we would expect lower bidding under the share treatment than under the probabilistic treatment. This is because, under the share treatment, there is no clear winner, and there are no probabilities to be distorted.

Information on the treatments is summarised in the following table:

| Sessions | # subjects | $n$ | P/S | L/C | RNNE bid | Observations |
|---|---|---|---|---|---|---|
| 1–3 | 12 | 4 | P | C | 15 | 1080 |
| 4–6 | 12 | 4 | P | L | 15 | 1080 |
| 7–9 | 12 | 4 | S | C | 15 | 1080 |
| 10–12 | 12 | 4 | S | L | 15 | 1080 |

Note that it is a 2 × 2 full-factorial design, with all possible combinations of the two treatments being covered. This means that it will be possible to estimate the interaction effect between the two treatments in addition to the main effects.

### *4.8.3 Analysis of data from contest experiment*

The data of Chowdhury et al. (2014) are contained in the file **chowdhury**. To give a feel for the data, we present a histogram of effort levels, for the 1,080 observations in the baseline treatment (PL), in Figure 4.4. It is seen that there is a large variation in effort levels, covering the entire permissible range of the variable. We also see that the distribution is multi-modal, with accumulation of effort levels at multiples of 5. Most importantly, we see a tendency for effort to exceed the Nash equilibrium prediction of 15 (i.e. a tendency for over-bidding). The mean effort level over the 1,080 observations in the baseline treatment is 26.2, implying that the mean over-bidding rate is 75%.

The analysis performed here is similar to that of Chowdhury et al. (2014). We define a variable $o_{it}$ to be the "over-bid", that is, the excess effort relative to the Nash equilibrium prediction, by subject $i$ in round $t$. This is simply:

$$o_{it} = e_{it} - 15 \tag{4.14}$$

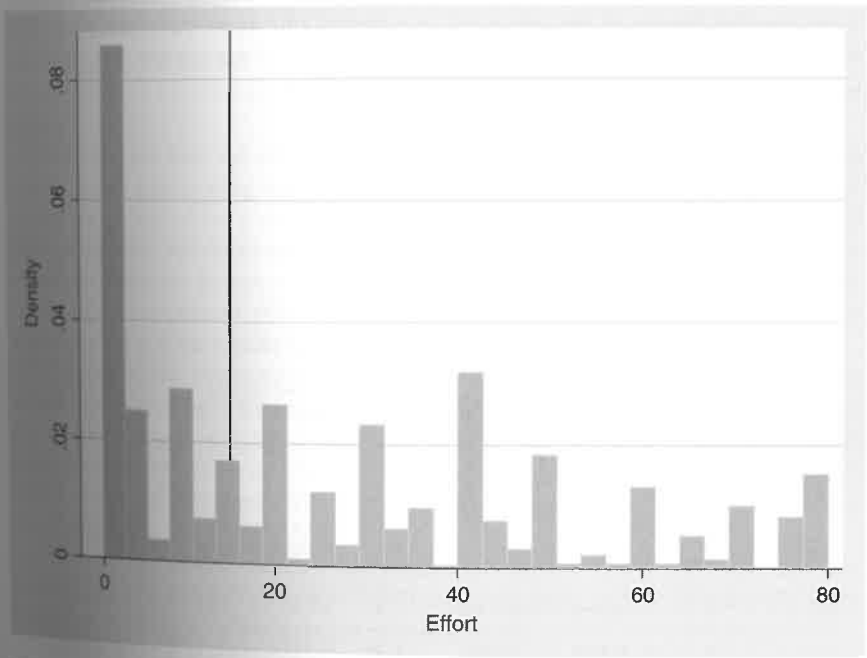



Figure 4.4: Distribution of effort levels in the experiment of Chowdhury et al.'s (2014). All 1,080 observations in the PL treatment. Vertical line drawn at Nash equilibrium prediction.

We then consider the following random effects model with $o_{it}$ as the dependent variable:

$$o_{it} = \beta_0 + \beta_1 S_i + \beta_2 C_i + \beta_3 S_i \star C_i + \beta_4 (1/t) + u_i + \epsilon_{it} \qquad (4.15)$$

$S_i$ and $C_i$ are treatment dummies for share and convex respectively, and $S_i \star C_i$ is the interaction variable formed as the product of the two. The intercept parameter $\beta_0$ has a clear interpretation as the expected level of over-bidding in the baseline treatment (PL) with experience (i.e. when $t$ is large).

Chowdhury et al. (2014) perform estimation separately by treatment, and also separately for the first 15 and last 15 rounds. The reason for estimating separately for the first 15 and last 15 rounds is that the (negative) effect of experience on effort appears to diminish markedly over the 30 rounds. They provide eight sets of estimation results in their Table 3. To reproduce their first set of results, we would use the following STATA code:

```
gen o = bid - 15
xtset i t
xtreg o s t if (c==1)&(t<=15)
```

When the command `xtreg` is used with no options, the random effects model is estimated as the default.

Here, we shall instead estimate model (4.15) using all data. By using all data we are able to include both treatments together, both with and without an interaction. Note also that the use of the reciprocal of $t$ instead of $t$ as an explanatory variable in (4.6) is a means of capturing the diminishing effect of experience uncovered by Chowdhury et al. (2014). The code we use is therefore:

```
gen o = bid - 15
xtset i t
gen sc=s*c
gen rec_t=1/t
xtreg o s c rec_t
xtreg o s c sc rec_t
```

The results are:

```
. *Random effects model without interaction:
. xtreg o s c  rec_t

Random-effects GLS regression                   Number of obs      =      4320
Group variable: i                               Number of groups   =       144

R-sq:  within  = 0.0223                         Obs per group: min =        30
       between = 0.0883                                        avg =      30.0
       overall = 0.0497                                        max =        30

                                                Wald chi2(3)       =    104.72
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           o |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -7.234259   1.986886    -3.64   0.000    -11.12848   -3.340034
           c |  -1.249074   1.986886    -0.63   0.530      -5.1433    2.645151
      rec_t |   11.51388   1.180853     9.75   0.000     9.199449    13.82831
      _cons |   11.64993   1.727864     6.74   0.000     8.263383    15.03648
------------------------------------------------------------------------------
```

```
    sigma_u |  11.614135
    sigma_e |  14.727186
        rho |  .38344663   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. *Random effects model with interaction:
. xtreg o s c sc rec_t

Random-effects GLS regression                   Number of obs      =      4320
Group variable: i                               Number of groups   =       144

R-sq:  within  = 0.0223                         Obs per group: min =        30
       between = 0.1150                                        avg =      30.0
       overall = 0.0608                                        max =        30

                                                Wald chi2(4)       =    113.26
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           o |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -3.194537   2.778254    -1.15   0.250    -8.639815    2.250741
           c |   2.790648   2.778254     1.00   0.315     -2.65463    8.235926
          sc |  -8.079444   3.929045    -2.06   0.040    -15.78023   -.3786583
       rec_t |   11.51388   1.180853     9.75   0.000     9.199449    13.82831
       _cons |   9.630073   1.970806     4.89   0.000     5.767365    13.49278
-------------+----------------------------------------------------------------
    sigma_u |  11.476361
    sigma_e |  14.727186
        rho |  .37781999   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Let us first interpret the results of the model without the interaction. The intercept is estimated to be +11.6 and the estimate is strongly significant. This simply tells us that in the baseline treatment (PL) with experience, subjects tend on average to over-bid by 11.6. This amounts to a rejection of the fundamental prediction of the theory.

We also see that the effect of experience is highly important. Effort declines rapidly but at a decreasing rate, as evidenced by the significantly positive coefficient on the reciprocal of the round number. Subjects appear to move in the direction of the equilibrium, but they converge to a point some distance short of it. In fact, the closeness of the coefficient of $1/t$ to the estimate of the intercept tells us that in the course of the experiment, subjects converge to a point roughly half way between the starting point and the equilibrium.

Turning to the treatment effects, we first see that share treatment significantly reduces over-bidding. This result is consistent with the "joy of winning" hypothesis; under the share rule, there is no clear winner, and hence any "joy of winning" motivation must be reduced. The result could also be explained in terms of the probability distortion hypothesis. The convex cost treatment appears to have no effect in the first model.

The second model includes the interaction variable "$s \star c$". We see that this variable has a significantly negative coefficient, indicating that the effect of the share treatment is stronger in the presence of convex costs.

Chowdhury et al. (2014) suggest that these findings regarding the drivers of out-of-equilibrium play are useful for the robust design of contests, and in particular that the interactions between these drivers may be important.

## 4.9 Meta-analysis

Meta-analysis is the name given to the set of methods for combining results from different studies, with the objective of identifying stronger patterns than can be seen in individual studies. In this section, we demonstrate that certain interesting experimental results (including comparative static predictions and effects of design features) can be confirmed (or refuted) using meta-analysis. We continue with the theme of experimental contests.

In the last section, data from a contest experiment conducted by Chowdhury et al. (2014) was analysed, and it was noted that the mean over-bidding rate in their baseline treatment was 75%. Sheremeta (2013) has collected a set of 39 such over-bidding rates (including that one) from experiments reported in 30 published articles. Along with the over-bidding rates, a number of features of each experiment are recorded, including number of players, prize, endowment, and matching protocol. The data set is presented in Table 1 of Sheremeta (2013), and it is reproduced in the file **sheremeta**. The first 20 rows of the data set are presented in Figure 4.5.

An obvious issue arising when meta-analysis is applied to contest results is that studies are not directly comparable in the sense that they use different units of measurement for the endowment and prize. To make the studies comparable, we therefore define the variable *endowment_rel* (relative endowment) as endowment divided by prize. Another issue is that, as seen in Figure 4.5, the data has an element of clustering, with some articles generating more than one row of the data. While this feature of meta-data may be very important in some situations, adjusting for clustering (using the techniques described earlier in the current chapter) makes very

| | obs | study | author | year | treatment | matching | endowm |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | millner_pratt | 1989 | lottery | random | |
| 2 | 2 | 2 | millner_pratt | 1991 | less RA | random | |
| 3 | 3 | 3 | shogren_baik | 1991 | lottery | fixed | |
| 4 | 4 | 4 | davis_reilly | 1998 | lottery | fixed | |
| 5 | 5 | 5 | potters_etal | 1998 | lottery | random | |
| 6 | 6 | 6 | anderson_stafford | 2003 | homogeneous | one-shot | |
| 7 | 7 | 6 | | . | | one-shot | |
| 8 | 8 | 6 | | . | | one-shot | |
| 9 | 9 | 6 | | . | | one-shot | |
| 10 | 10 | 6 | | . | | one-shot | |
| 11 | 11 | 7 | schmitt_etal | 2004 | static | random | |
| 12 | 12 | 8 | schmitt_etal | 2005 | single-prize | one-shot | |
| 13 | 13 | 9 | herrmann_orzen | 2008 | direct repeated | random | |
| 14 | 14 | 10 | kong | 2008 | less RA | fixed | |
| 15 | 15 | 11 | fonseca | 2009 | simultaneous | random | |
| 16 | 16 | 12 | abbink_etal | 2010 | one:one | fixed | |
| 17 | 17 | 13 | sheremeta | 2010 | one-stage | random | |
| 18 | 18 | 14 | sheremeta_zhang | 2010 | individual | random | |
| 19 | 19 | 15 | ahn_etal | 2011 | individual | fixed | |
| 20 | 20 | 16 | deck_jahedi | 2011 | baseline | one-shot | |

Figure 4.5: First 21 rows of the data set of Sheremeta's (2013)

little difference to the analysis that follows. For this reason the clustering will be disregarded.

Equation (2) of Sheremeta (2013) presents the results of a linear regression with the over-bidding rate as the dependent variable and four explanatory variables: relative endowment; number of players (n); partners matching dummy; and one-shot dummy (the strangers matching dummy is excluded and represents the base case). The results are exactly reproduced as follows.

```
. reg overbid endowment_rel n partners one_shot

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  4,    34) =    8.73
       Model |  6.49674324     4  1.62418581            Prob > F      =  0.0001
    Residual |  6.32637945    34  .186069984           R-squared     =  0.5066
-------------+------------------------------           Adj R-squared =  0.4486
       Total |  12.8231227    38  .337450597           Root MSE      =  .43136

------------------------------------------------------------------------------
     overbid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
endowment_rel |    .431265   .2061099     2.09   0.044     .0123993    .8501307
           n |    .2036022   .0414801     4.91   0.000     .1193046    .2878999
    partners |   -.0778284   .1690991    -0.46   0.648    -.4214791    .2658223
    one_shot |    .2929277   .1984659     1.48   0.149    -.1104035    .6962588
       _cons |   -.4108494   .2713689    -1.51   0.139    -.9623374    .1406386
------------------------------------------------------------------------------
```

The $R^2$ of the regression is slightly more than 0.50, indicating that more than half of the variation in over-bidding rates is being explained by this simple model. The results also show that the over-bidding rate rises with the relative endowment ($p < 0.05$) and also rises with the number of players ($p < 0.01$). The two other coefficients indicate that the matching protocol is not important in explaining over-bidding rates. It is conceivable that "partners" matching might result in lower effort as a consequence of collusion between players, but this effect is not seen in this data. Baik et al. (2014) have already found evidence of this "non-result" in the context of a single experiment in which matching protocol varies between treatments. They stress the usefulness of the "non-result" in the sense that it implies that partners matching can reasonably be used in preference to strangers matching, with its attendant advantages including the greater number of independent observations for a given total number of subjects.

Let us investigate the effect of relative endowment more closely. In Figure 4.6, we present a scatter plot of over-bidding rates against relative endowment. Superimposed on the scatter plot is a "Lowess smoother", a form of non-parametric regression first developed by Cleveland (1979). This indicates that the relationship between endowment and effort may be non-linear, and in particular that there may be an "optimal" (from the viewpoint of the contest organiser at least!) relative endowment, around one, at which effort is maximised. Baik et al. (2014) have already found evidence of this result in the context of a single experiment in which endowment is a treatment variable. They attribute this result to the endowment acting as a constraint when it is low, but generating a wealth effect when it is larger. Wealth is hypothesised to bring about a reduction in "conflict intensity", that is, a reduction in effort.

| treatment | matching | endowment |
|---|---|---|
| lottery | random | |
| less RA | random | |
| lottery | fixed | |
| lottery | fixed | |
| lottery | random | |
| homogeneous | one-shot | |
| | one-shot | |
| | one-shot | |
| | one-shot | |
| | one-shot | |
| static | random | |
| single-prize | one-shot | |
| Direct repeated | random | |
| less RA | fixed | |
| simultaneous | random | |
| one:one | fixed | |
| one-stage | random | |
| individual | random | |
| individual | fixed | |
| baseline | one-shot | |

of Sheremeta's (2013)

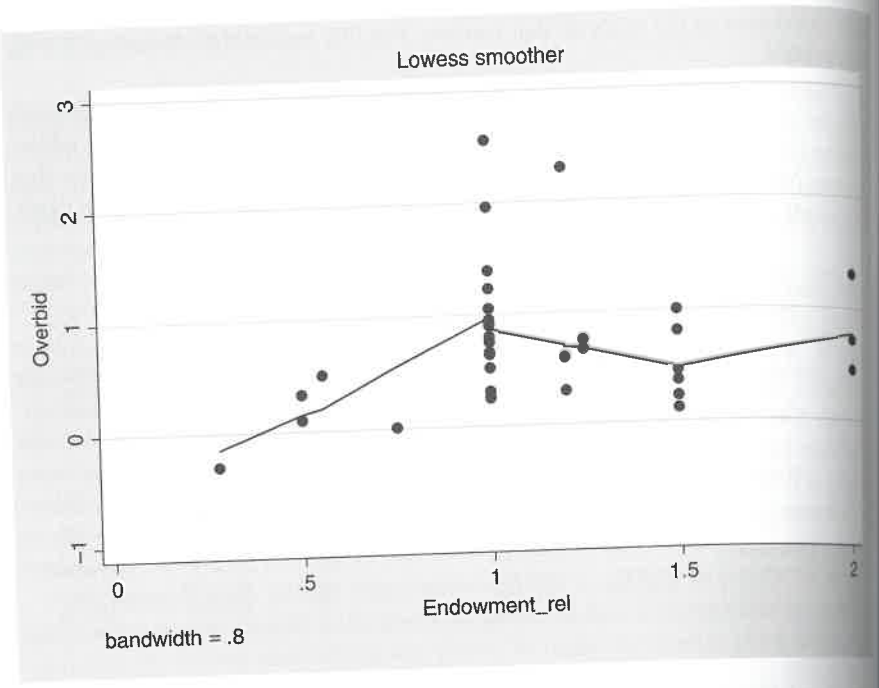



Figure 4.6: Scatterplot and Lowess smoother of over-bidding rate against relative endowment

The non-linear effect of endowment may be tested in the context of the meta-analysis by including the square of relative endowment (`end2` in the data set) as an explanatory variable. The results are as follows:

```
. reg  overbid endowment_rel end2 n partners one_shot

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  5,    33) =    9.86
       Model |  7.68145525     5  1.53629105            Prob > F      =  0.0000
    Residual |  5.14166744    33  .155808104            R-squared     =  0.5990
-------------+------------------------------           Adj R-squared =  0.5383
       Total |  12.8231227    38  .337450597            Root MSE      =  .39473

------------------------------------------------------------------------------
     overbid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
endowment_rel |    2.37303   .729003     3.26   0.003     .8898622    3.856198
        end2 |   -.8146338   .2954276   -2.76   0.009    -1.415686   -.2135816
           n |    .1986473   .0379999    5.23   0.000     .1213359    .2759587
    partners |    .0185187   .1586342    0.12   0.908     -.304225    .3412624
    one_shot |    .3408628   .1824413    1.87   0.071    -.0303169    .7120424
       _cons |   -1.471736   .4579111   -3.21   0.003    -2.403363   -.5401089
------------------------------------------------------------------------------
```

We see that the coefficients of relative endowments and its square are respectively strongly positive and strongly negative, implying that there is indeed a value of relative endowment at which effort is maximised, confirming the result of Baik et al. (2014). It is easily verified that this optimal level can be computed as $-\frac{\hat{\beta}_1}{2\hat{\beta}_2}$, where

$\hat{\beta}_1$ and $\hat{\beta}_2$ are respectively the coefficients on relative endowment and its square. This computation can be performed in STATA using a technique known as the *delta method*, which will be explained in detail in Chapter 6. The STATA command is `nlcom` and this is required immediately after the `regress` command. The command along with the results are as follows:

```
. nlcom end_star: -_b[endowment_rel]/(2*_b[end2])

   end_star:  -_b[endowment_rel]/(2*_b[end2])

------------------------------------------------------------------------------
     overbid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    end_star |   1.456501   .1503841     9.69   0.000     1.150542     1.76246
------------------------------------------------------------------------------
```

We see that the estimate of the "optimal endowment" is 1.46. A major attraction of using the delta method is that it also returns a standard error and 95% confidence interval. The confidence interval is [1.15, 1.76] which conveys evidence that the optimal relative endowment is somewhat greater than one.

## 4.10 Summary and Further Reading

In this chapter, we have tried to provide an overview of the types of methods used in the analysis of data from experiments in which subjects are required to submit bids. This set of techniques is applicable over a wide range of experimental settings. The settings on which we have focussed for illustrative purposes are auction experiments and contest experiments.

Readers wishing to learn about auction theory are referred to Krishna (2010). Studies in which auction data is modelled using the type of techniques introduced in this chapter include Kagel et al. (1995), Kagel & Levin (1986) and Ham et al. (2005).

Clustering has been a recurring theme. For a useful recent dicussion of the importance of clustering in economic experiments, see Fréchette (2012). For panel data estimation techniques, readers are referred to Baltagi (2008), and for multi-level modelling, they are referred to Rabe-Hesketh & Skrondal (2008).

Drichoutis et al. (2011) apply a dynamic panel estimation procedure to allow for bid interdependence (between rounds) in auctions. This sort of estimation procedure allows (for example) the bid in the previous round to influence the bid in the current round. Readers interested in dynamic panel data models should consult Roodman (2009). These estimation techniques have not been covered in this chapter, although the topic of dynamic panel estimation will be covered briefly in Section 9.3.3.

The technique of meta-analysis was briefly introduced in this chapter, and the meta-analytic results of Sheremeta (2013) in the context of contest experiments were reproduced and extended. Other meta-analyses that have been published in the experimental economics literature include Zelmer (2003) (public goods games), Engel (2011) (dictator games), and Johnson & Mislin (2011) (trust games).

## Exercises

1. Following each of the regression models estimated in Sections 4.5 and 4.6, the RNNE theory was tested using Wald tests. Conduct tests of the same hypothesis using LR tests. Do the results agree?
2. Consider the expected payoff for player $i$ in a Tullock contest with $n$ players, as given in (4.11). Substitute the contest success function (4.10) into (4.11), differentiate with respect to $e_i$, set to zero, and finally assume that all $n$ players invest the same effort, in order to find the Nash equilibrium effort level defined by (4.12).

l in Sections 4.5 and 4.6, the
t tests of the same hypothesis

llock contest with $n$ players,
function (4.10) into (4.11),
ally assume that all $n$ players
uilibrium effort level defined

# Chapter 5
# Modelling of Decision Times Using Regression Analysis

## 5.1 Introduction

In this chapter we consider the application of regression analysis to a topic that is currently growing in importance: the modelling of decision times. The decision time applied by a subject to a given task is often measured electronically with great accuracy, and is a reliable measure of the effort expended in performing the task. This sort of analysis is useful for a number of reasons, most importantly that it enables us to identify the features of a task that tend to increase the effort expended by subjects.

The decision-time example is also useful in the further demonstration of panel data estimators and in the highlighting of differences between estimation techniques. The example we use is particularly useful because all of the explanatory variables appearing in the model are time-varying. This means that, unlike with the auctions data in Chapter 4, the fixed-effects estimator can be used, and the Hausman test may be demonstrated as a test for adjudicating between fixed and random effects.

The example we consider in this chapter is decision times in a (simulated) risky choice experiment: linear regression models are used to identify the determinants of the length of time taken to choose between two lotteries. The results will be interpreted in terms of subjects' allocation of cognitive effort. A similar analysis has been performed by Moffatt (2005*b*). Camerer & Hogarth's (1999) "capital-labour-production" framework is relevant and useful here. "Capital" is the knowledge, skills, and experience which the subject brings to the experiment, and also includes the experience acquired during the experiment. "Labour" is the mental effort exerted while solving a task. "Production" is loosely represented by the subject's performance in the task, and is obviously determined by the levels of capital and labour inputs. Although there are reasons for believing that capital input is fixed at least in the context of a single experiment, labour input is fully flexible and potentially highly responsive to incentives and other factors.

Capital input as defined above is clearly hard to measure accurately. However, we shall see that the effects of knowledge and experience are seen indirectly on more than one occasion in estimation. In contrast, a measure of "labour input" is readily available: the time in seconds taken to make each choice.

A number of factors might be expected to influence labour input (i.e. effort). These will be introduced with the aid of a motivating example in Section 5.2. In Section 5.3, we develop a theoretical model of effort allocation, with a view to testing it in the later empirical analysis. Section 5.4 presents the data and considers estimation of the effort model using linear regression. Section 5.5 progesses to panel data estimation of the same model, and includes the demonstration of the Hausman test applied to adjudicating between random effects and fixed effects. Results are discussed in Section 5.6. Section 5.7 considers post-estimation issues, in particular a method for extracting the "posterior random effect", and why this is useful. Section 5.8 provides a summary.

## 5.2 The Decision-time Data

The data used in this chapter is simulated, and is from the same (simulated) experiment as used in Chapter 13. It is contained in the file **decision_times_sim**. The data is simulated in such a way as to resemble a real data set as closely as possible. This was achieved by basing the simulation, of both choices and decision times, on the results of Moffatt (2005*b*), who analysed the real data set of Hey (2001). Full details of the simulation can be found in Chapter 13.

The simulation is of the lottery choices of an imaginary sample of 60 subjects, each of whom faced a set of 50 choice problems on two different days. The total number of observations in the data set is therefore $60 \times 100 = 6{,}000$. The ordering of the problems changed between sessions and also between subjects. The probabilities defining the 50 choice problems are listed in Appendix C. All 50 problems involve the three outcomes: \$0, \$10, and \$20. We imagine that the random lottery incentive system was applied: at the end of the second session, one of the subject's 100 chosen lotteries was selected at random and played for real.

We assume that, for each choice made, decision time was electronically recorded. There are therefore 6,000 decision times in the sample. The penultimate column of the table in Appendix C shows the mean decision time (in seconds) for each choice problem.

Direct motivation for the analysis of decision times reported later is provided by briefly previewing two examples of risky choice problems from the experiment. The selected problems are presented diagrammatically, as they would be presented to subjects during the experiment (if it were real), in Figure 5.1. The mean of decision time is taken over the 120 observed decisions for each problem, and is shown in the figure.

The key feature of the example is that "Task 40" takes subjects almost three times as long to solve, on average, as "Task 14". There are many possible explanations for the difference. The most obvious is differing complexity: the second problem is clearly more complex than the first. This is not the most interesting determinant of effort, but it is clearly a factor that needs to be controlled for. An important issue to be addressed later is therefore how best to measure complexity. A second possible explanation of the difference in decision times is the difference

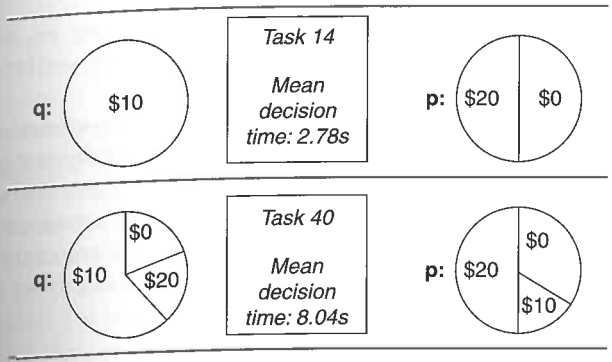



Figure 5.1: Typical choice tasks in the risky-choice experiment

in financial incentives: the expected values of the lotteries are higher in the second problem. A third, and more subtle, explanation for the difference is the possibility that subjects tend to be closer to indifference when faced with the second problem, than with the first. In order to investigate this possibility, we need to establish a framework for measuring closeness to indifference for each subject and for each problem. This framework is the parametric choice model developed and estimated in Chapter 13. A variable representing closeness to indifference is obtained using the method described in Section 13.4.5. This variable is used in the modelling in this chapter.

## 5.3 A Theoretical Model of Effort Allocation

In this section, we adopt a framework introduced by Moffatt (2005*b*) to analyse the effects of two of the factors that we expect to be important in the allocation of effort: the subject's closeness to indifference between the two lotteries; and the objective similarity between the two lotteries. It is important to realise that these two factors are not equivalent. If two lotteries are identical, then necessarily a subject will be indifferent between them, but the converse does not apply: indifference does not imply identicality. In order to isolate the effect of closeness to indifference, we therefore need to control for the effect of objective similarity.

As noted in the previous section, all choice problems in the experiment involve combinations of the three outcomes: \$0, \$10, \$20. We index the problems in the experiment by $t$ ($t = 1, \ldots, 100$). For most choice problems, 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 problem is one of "dominance", since one lottery first-order stochastically dominates the other.[1] If task $t$ is a non-dominance problem, we will label the riskier lottery as $\mathbf{p}_t$, and the safer as $\mathbf{q}_t$. For a dominance problem, $\mathbf{p}_t$

[1] See Chapter 12 for definitions of these terms.

will be the dominating lottery, and $\mathbf{q}_t$ the dominated. $\mathbf{p}_t = (p_{1t}\ p_{2t}\ p_{3t})$ and $\mathbf{q}_t = (q_{1t}\ q_{2t}\ q_{3t})$ are vectors containing three probabilities corresponding to the three possible outcomes.

For closeness to indifference, we shall use the absolute valuation differential for subject $i$ in problem $t$, $|\hat{\Delta}_{it}|$. This variable is explained fully, and generated, in Chapter 13. In short, it is a non-negative variable which takes the value zero if subject $i$ is completely indifferent between the two lotteries in problem $t$, and takes a large positive value if the subject has a clear preference for one of the lotteries.

For objective similarity (of the lotteries in choice problem $t$) we shall use the following measure:

$$\Delta_t^o = \sum_{j=1}^{3} (q_{jt} - p_{jt})^2 \tag{5.1}$$

Note that $\Delta_t^o = 0$ for a problem in which the two lotteries $\mathbf{p}_t$ and $\mathbf{q}_t$ are identical, while $\Delta_t^o$ reaches a maximum of 2 if the two lotteries are certainties of different amounts.

Figure 5.2 shows a graph with a particular subject's absolute valuation differential on the horizontal, and objective difference on the vertical. First, we note that the feasible region is the triangle OAC. This is because when the two lotteries are identical, the subject is necessarily indifferent, so we must be at the origin. Also, when the objective difference is maximal, that is, when the two lotteries are certainties of different amounts, it is not possible for the subjective valuation differential to be zero, hence the point B is outside the feasible region. The upward sloping lines

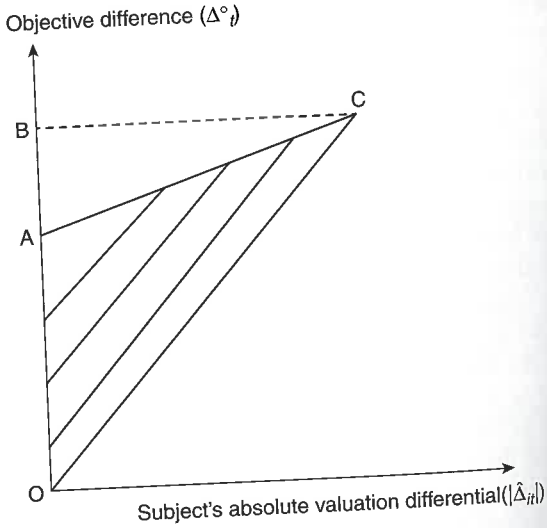



Figure 5.2: Iso-effort curves in $\left(|\hat{\Delta}_{it}|, \Delta_t^o\right)$ space

inside the triangle, including OC, can be interpreted as *iso-effort* curves. Along OC, effort is minimised, since here the subject's preference is as clear as is allowed for a pair of lotteries which differ objectively by a given amount. The other *iso-effort* curves represent higher levels of effort, and the highest effort is allocated at point A, where although the lotteries are objectively very different, the subject in question is indifferent between them.

The important predictions arising from the simple analysis depicted in Figure 5.2 are that, *ceteris paribus*, we expect effort to increase with the objective difference between the lotteries, but to decrease with the subject's absolute valuation differential. The econometric models of effort allocation reported in Sections 5.4–5.6 confirm these predictions, albeit on simulated data.

## 5.4 An Econometric Model of Effort Allocation

As mentioned in Section 5.1, our measure of "labour input", or effort expended in solving a problem, is simply the time taken to make a decision. The logarithm of this variable will be the dependent variable in the model we will estimate in this section.

To give a feel for the data, summary statistics of pooled decision times are shown in Table 5.1, and a histogram of the same variable is shown in Figure 5.3. We see that a typical response time is between 0 and 10 seconds, although there is a very long tail to the right. This calls for the use of the logarithmic transformation of the variable in econometric modelling. Multiplying the mean by 50, we see that a typical subject would spend around five minutes on each of the two days engaging in the experiment.

Like many experimental data sets, this one contains repeated observations for each subject. This should be taken into account in modelling. As explained in Chapter 4, it is natural to handle such data sets in a panel data framework, in which it is assumed that, say, $n$ subjects are observed making a decision in each of $T$ time periods. Clearly, subjects are expected to differ from each other. In the present context, it is expected that some subjects are by nature quick decision makers, and that others are slow. This clearly implies that, if we are treating decision time as the dependent variable in a regression model, we anticipate dependence at the level of the subject, and we need to deal with this dependence using the methods demonstrated in a different context in Chapter 4.

A very useful way of viewing panel data graphically is using the `xtline` command in STATA. First, the `xtset` command is used to declare the data set as a panel.

| Variable | n | mean | median | s.d. | min | max |
|---|---|---|---|---|---|---|
| Decision time (seconds) | 6000 | 5.098 | 3.808 | 4.6235 | 0.231 | 73.02 |

Table 5.1: Summary statistics of decision time in seconds

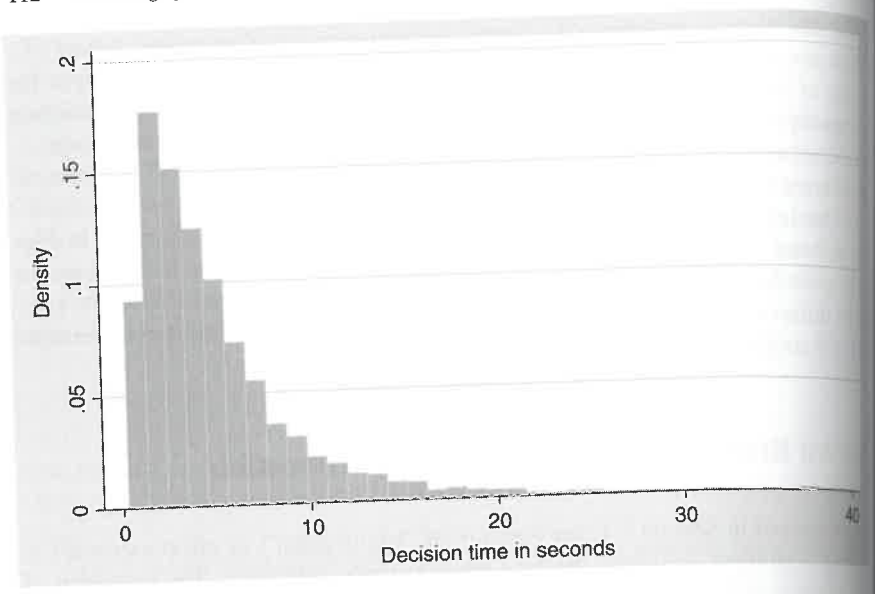



Figure 5.3: Histogram of decision time

Then we apply the `xtline` command to decision time, excluding the small number of observations for which decision time is greater than 20 seconds.

```
. xtset i t
      panel variable:  i (strongly balanced)
       time variable:  t, 1 to 100
               delta:  1 unit

. xtline dt if dt<20
```

The result is shown in Figure 5.4. We see that the `xtline` command produces a time series plot of the decision time variable for each of the 60 subjects in the sample. The resulting graph is useful for assessing the extent of between-subject variability. For some subjects, the time series is persistently very close to zero, indicating fast decision times. Subject 24 is in fact the fastest decision maker, with a mean decision time of 1.558 seconds.[2] For other subjects, decision time is persistently high, indicating slow decision making. The slowest decision maker appears to be subject 29, with a mean decision time of 9.947 seconds.

Next, we shall use graphical analysis to identify the determinants of decision time. In Figure 5.5 the logarithm of decision time is plotted against a variable representing the position of the problem in the sequence of 100 problems. Since it is hard to discern a relationship from the scatter alone, a non-parametric regression (Lowess; see Cleveland, 1979), sometimes called a "smoother", has been

[2] The mean of decision time for individual subjects are computed using the command:

```
table i, contents (mean dt)
```


Figure 5.4: Time series graphs of decision time (in seconds) by subject

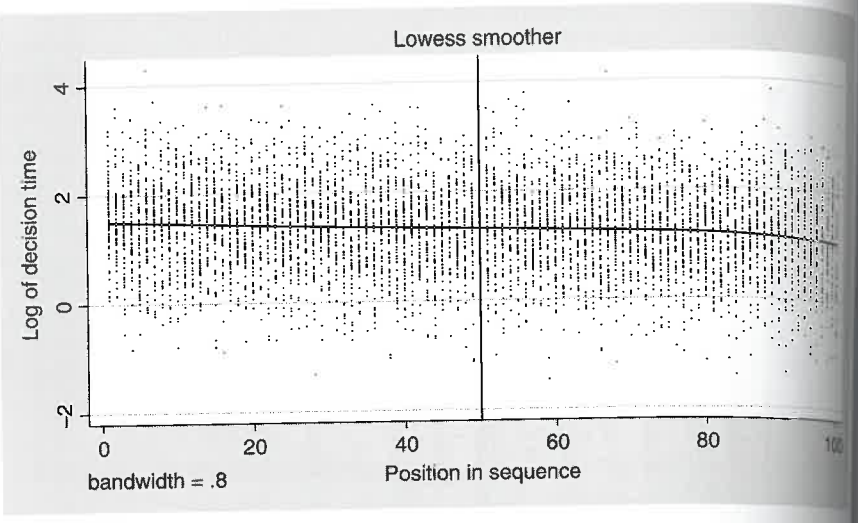



Figure 5.5: Logarithm of decision time against position in the sequence

superimposed. This smoother clearly reveals a tendency for decision times to diminish with the progress of the experiment. Also, remembering that 50 problems were solved on each day, a within-day reduction in decision times is also discernible in Figure 5.5 (this is seen most easily in the extremes of the distribution). While the overall decline over the experiment may be attributed to the accumulation of experience, any within-day decline is more likely to be the result of boredom. These two effects are captured separately in the model estimated in the next section.

In Figure 5.6 decision time in seconds is plotted against the absolute valuation differential (closeness-to-indifference), again with a smoother. The smoother clearly reveals that more effort is allocated to problems for which the subject is close to indifference. This relationship will be confirmed in the estimation results of the parametric model.

As mentioned in Section 5.2, it is very likely that effort expended depends upon the complexity of the problem, and therefore we must control for this in some way. We assume that subjects assess the complexity of a problem using a very simple rule: they count the number of outcomes appearing in the simpler of the two lotteries.[3] This rule gives rise to three levels of complexity, to which we shall refer as Level 1, Level 2, and Level 3. For specific examples, we may refer back to Figure 5.1, and note that task 14 is of complexity level 1, since one lottery involves only a single outcome of \$10, while task 40 is of the higher level of complexity 3, since both lotteries involve this number of outcomes.

[3] A finer classification of problems was originally used, but the estimated model indicated that the simple classification into three levels is sufficient to explain the data. Hey (1995) and others have used a different measure: the mean taken over the two lotteries of the number of outcomes.

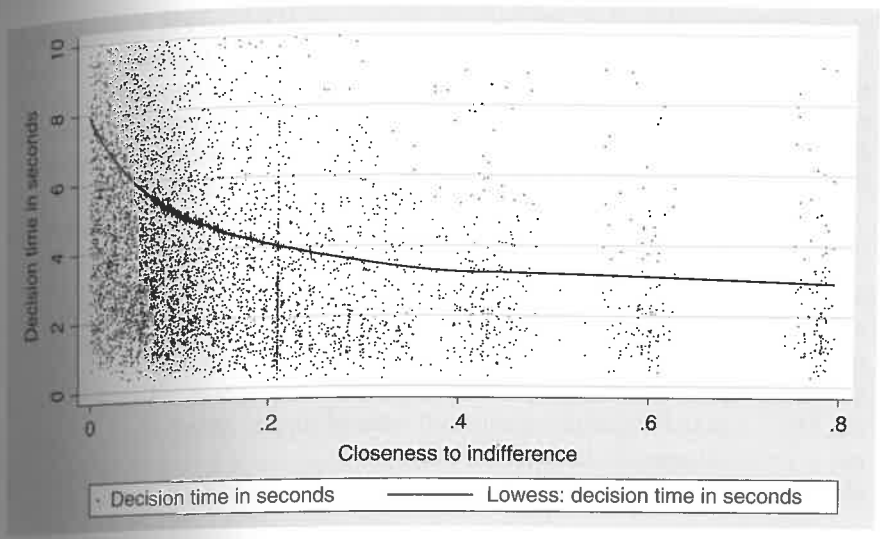



Figure 5.6: Decision time against closeness to indifference

In the following section, we will estimate the following (log-)linear regression model, with the logarithm of decision time as the dependent variable:

$$\begin{aligned} \log(decision\ time_{it}) &= \alpha + \beta_1 complex2_t + \beta_2 complex3_t + \beta_3 \tau_{it}^d + \beta_4 \tau_{it} \\ &+ \beta_5 log(EV_t) + \beta_6 |\hat{\Delta}_{it}| + \beta_7 |\hat{\Delta}_{it}|^2 + \beta_8 |\hat{\Delta}_{it}|^3 + \beta_9 \Delta_t^o + u_i + \epsilon_{it} \qquad (5.2) \\ & i = 1 \ldots, n \quad t = 1 \ldots, T \quad var(u_i) = \sigma_u^2 \quad var(\epsilon_{it}) = \sigma_\epsilon^2 \end{aligned}$$

In (5.2), $i$ indexes subjects, while $t$ indexes problems. Note that there are two stochastic terms: $u_i$ is the subject-specific effect, with mean zero and variance $\sigma_u^2$; $\epsilon_{it}$ is the equation error, with mean zero and variance $\sigma_\epsilon^2$. Equation (5.2) is either a fixed effects (FE) or a random effects (RE) model, depending on how the individual effect $u_i$ is interpreted. Both of these models will be estimated in the next section. We shall also estimate a model in which $u_i$ is assumed to be zero for all $i$. Such a model is known as a pooled regression model, since it disregards the panel structure of the data.

The first two explanatory variables in (5.2) are dummy variables indicating the level of complexity of problem $t$, according to the rule defined above. The excluded complexity level is the least complex, Level 1. The third and fourth explanatory variables represent the position of a problem in the sequence: $\tau_{it}^d$ is the position of problem $t$ within the day on which the problem was solved, so $\tau^d$ ranges from 1 through 50; $\tau_{it}$ is in contrast the position of problem $t$ in the complete sequence of problems faced by subject $i$, and therefore ranges from 1 through 100. The fifth explanatory variable represents the financial incentive associated with each problem. Many different measures could be used here; the one which is chosen is the logarithm of the expected value of the simpler of the two lotteries. The next three

explanatory variables are the closeness to indifference of subject $i$ in choice problem $t$, as defined briefly in Section 5.3, and its square and cube. The purpose of these three variables is to allow closeness to indifference to have a non-linear effect on effort, as it appears to do on the evidence of the non-parametric regression shown in Figure 5.6. The final explanatory variable is our measure of the objective difference between the two lotteries, as defined in (5.1).

The various variables in the data set (**decision_times_sim**) are named as follows:

| | |
|---|---|
| log_dt: | natural logarithm of decision time in seconds; |
| complex: | level of complexity of choice problem (1,2, or 3); |
| tau_d: | position of choice problem within single day (1–50) ($\tau_{it}^{d}$); |
| tau: | position of choice problem in complete sequence of 100 ($\tau_{it}$); |
| log_ev: | natural logarithm of expected value of simpler lottery; |
| cti: | closeness-to-indifference ($\lvert\hat{\Delta}_t\rvert$); |
| obj_diff: | objective difference between the two lotteries ($\Delta_t^o$). |

## 5.5 Panel Data Models of Effort Allocation

We start by estimating (5.2) using a pooled model, that is, a regression using the complete sample of $nT$ observations, that attributes all unexplained variation in terms of within-subject randomness, and does not allow for any variation between subjects. In the context of (5.2), the pooled model is one in which $u_i = 0$ for all $i$. The following STATA command performs the required regression:

```
. regress  log_dt complex2 complex3 tau_d tau  logev cti cti2 cti3 obj_diff

      Source |       SS       df       MS              Number of obs =    6000
-------------+------------------------------           F(  9,  5990) =  116.66
       Model |  576.37795      9  64.0419944           Prob > F      =  0.0000
    Residual |  3288.28327  5990  .548962149           R-squared     =  0.1493
-------------+------------------------------           Adj R-squared =  0.1479
       Total |  3864.66122  5999  .644217573           Root MSE      =  .74093

------------------------------------------------------------------------------
      log_dt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    complex2 |   .2397661   .0493898     4.85   0.000     .1429443    .336588
    complex3 |   .3804073   .0654843     5.81   0.000     .2520345    .508780
       tau_d |   -.002203   .0007657    -2.88   0.004    -.0037041   -.000702
         tau |  -.0032846   .0003826    -8.58   0.000    -.0040346   -.002534
       logev |   .0609713   .0562583     1.08   0.279    -.0493151    .171258
         cti |  -5.590682   .3889394   -14.37   0.000    -6.353144   -4.828221
        cti2 |   12.31351   1.443477     8.53   0.000     9.483773    15.14325
        cti3 |  -8.576194   1.349663    -6.35   0.000    -11.22202   -5.930369
    obj_diff |   .1323685   .0428824     3.09   0.002     .0483035    .216433
       _cons |   1.648698   .0793198    20.79   0.000     1.493202    1.804193
------------------------------------------------------------------------------
```

We see that nearly all of the explanatory variables have strongly significant effects on decision time. However, we shall avoid interpreting each individual effect at this stage. This is because we will shortly be reporting estimates from panel data models

ence of subject $i$ in choice proble
e and cube. The purpose of the
nce to have a non-linear effect
on-parametric regression shown
neasure of the objective differen

**cision_times_sim**) are named

ı seconds;
m (1,2, or 3);
ngle day (1–50) ($\tau_{it}^d$);
lete sequence of 100 ($\tau_{it}$);
of simpler lottery;

ɔ lotteries ($\Delta_t^o$).

## location

del, that is, a regression using
butes all unexplained variation
ot allow for any variation betwe
el is one in which $u_i = 0$ for all
quired regression:

```
logev cti cti2 cti3 obj_diff

      Number of obs =
      F(  9,  5990) =
      Prob > F      =
      R-squared     =
      Adj R-squared =
      Root MSE      =

------------------------------
 P>|t|     [95% Conf. Interval]
------------------------------
 0.000      .1429443
 0.000      .2520345
 0.004     -.0037041
 0.000     -.0040346
 0.279     -.0493151
 0.000     -6.353144
 0.000      9.483773
 0.000     -11.22202
 0.002      .0483035
 0.000      1.493202
------------------------------
```

les have strongly significant effe
preting each individual effect a
ıg estimates from panel data mod



that are found to be superior to this pooled model. We shall interpret the results of the most preferred model in the next section.

On the evidence of Figure 5.4, we have reason to believe that there is strong dependence at the subject level, or "subject-level clustering". As in Chapter 4, we shall start to address this problem by simply correcting the standard errors for this sort of dependence. Recall that this requires using the `vce(cluster i)` option with the regress command.

```
regress log_dt complex2 complex3 tau_d tau logev cti cti2 cti3 obj_diff ///
  vce(cluster i)

Linear regression                               Number of obs =      6000
                                                F(  9,    59) =    196.50
                                                Prob > F      =    0.0000
                                                R-squared     =    0.1491
                                                Root MSE      =    .74092

                                    (Std. Err. adjusted for 60 clusters in i)
------------------------------------------------------------------------------
             |               Robust
      log_dt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    complex2 |   .2397661   .0395952     6.06   0.000     .1605363     .318996
    complex3 |   .3804073   .0529175     7.19   0.000     .2745197     .486295
      tau_d  |   -.002203   .0008415    -2.62   0.011    -.0038868    -.0005192
         tau |  -.0032846    .000355    -9.25   0.000    -.0039949    -.0025743
       logev |   .0609713    .047967     1.27   0.209    -.0350105     .1569531
         cti |  -5.590682   .3573599   -15.64   0.000    -6.305758    -4.875607
        cti2 |   12.31351   1.283165     9.60   0.000     9.745901     14.88111
        cti3 |  -8.576194   1.178483    -7.28   0.000    -10.93433    -6.218055
    obj_diff |   .1323685   .0340484     3.89   0.000     .0642379     .2004992
       _cons |   1.648698   .0713508    23.11   0.000     1.505925     1.79147
------------------------------------------------------------------------------
```

Note that the estimates themselves are identical to those obtained in the previous regression in which no adjustment was made for clustering. Only the standard errors have changed. Nearly all of the corrected standard errors are smaller than their uncorrected counterparts. This means that the t-statistics are larger in magnitude, and therefore greater significance is detected, as a result of making the correction. However, in this case, none of the changes are great enough to result in reversals of conclusions regarding significance of effects.

Using cluster-robust standard errors is only a first step in addressing the issue of the panel structure of the data. It is also possible to improve the estimates themselves, using a panel-data estimator instead of OLS. As discussed in Section 4.6 the two most popular panel data estimators are the fixed-effects and random-effects estimators. Both are represented by equation (5.2). Recall that in (5.2) there are two error terms: $\epsilon_{it}$ is the conventional equation error term, which is assumed to have mean zero and variance $\sigma_\epsilon^2$; $u_i$ is known as the subject-specific term. $u_i$ differs between subjects – hence the $i$-subscript – but it is fixed for a given subject. The two estimators differ in the way this term is interpreted.

The difference between fixed effects and random effects was explained in Section 4.6, and will be explained only briefly here. The fixed effects estimator is essentially a linear regression which includes a set of $n - 1$ dummy variables, one for each subject in the data set (with one excluded to avoid the dummy variable trap). Hence a different intercept is estimated for each subject. The random

effects estimator does *not* estimate the intercept for each subject. It simply recognises that they are all different, and sets out to estimate only their *variance*, $\sigma_u^2$. Note that random effects is more efficient than fixed effects, because there are far fewer parameters to estimate. We therefore prefer to use random effects if this model turns out to be acceptable.

The first question we might wish to ask is: is a panel data model required at all? If there are no differences between subjects, the pooled regression model estimated above is the correct model. The most obvious way of testing for differences within subjects is to test for equality of the subject fixed effects in the fixed-effects model. The null hypothesis for this test is:

$$H_0 : u_1 = u_2 = \cdots = u_n = 0$$

This null hypothesis embodies $n - 1$ restrictions on the model. The test is routinely performed as an F-test in the estimation of the fixed-effects model. If this null hypothesis is rejected (it nearly always is) we conclude that there are significant differences between subjects which make the pooled regression model invalid and necessitate the use of panel data estimation.

To decide between fixed effects (FE) and random effects (RE), the Hausman test is used. This test is based on a comparison of the two sets of estimates. The test will be explained in greater detail in Section 7.6. Roughly expressed the reasoning is as follows. The assumptions underlying RE are more stringent than those underlying FE. If the two sets of estimates are close to each other, this implies that the assumptions underlying both sets of estimates are valid, and RE is preferred because it is a more efficient estimator. If the two sets of estimates are very different from each other, this implies that only FE can be correct, and the assumptions underlying RE must be false. Hence, if the Hausman test results in a rejection, FE is preferred, while a failure to reject leads us to favour RE.

As mentioned in Section 4.6 panel data commands in STATA always start with the prefix `xt`. For example panel data (linear) regression is carried out using `xtreg`. The fixed effects and random effects estimators are implemented using the command with the options `fe` and `re` respectively.

The two models are estimated, and the Hausman test performed, with the following sequence of commands:

```
. xtset i t
       panel variable:  i (strongly balanced)
        time variable:  t, 1 to 100
                delta:  1 unit

. xtreg log_dt complex2 complex3 tau_d tau logev cti cti2 cti3 obj_diff, fe

Fixed-effects (within) regression               Number of obs      =
Group variable: i                               Number of groups   =

R-sq:  within  = 0.2004                         Obs per group: min =
       between = 0.0002                                        avg =
       overall = 0.1491                                        max =

                                                F(9,5931)          =
corr(u_i, Xb)  = -0.0059                        Prob > F           =
```

```
------------------------------------------------------------------------------
      log_dt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    complex2 |   .2374196   .0417019     5.69   0.000     .1556687    .3191706
    complex3 |   .3770243   .0552954     6.82   0.000     .2686252    .4854234
      tau_d  |  -.0022045   .0006465    -3.41   0.001    -.0034718   -.0009372
         tau |  -.0032846    .000323   -10.17   0.000    -.0039179   -.0026513
       logev |   .0601018   .0474991     1.27   0.206    -.0330137    .1532173
         cti |  -5.671949    .330508   -17.16   0.000    -6.319865   -5.024033
        cti2 |   12.46305   1.22361     10.19   0.000     10.06433    14.86177
        cti3 |  -8.645294   1.142589    -7.57   0.000    -10.88518   -6.405403
    obj_diff |   .1324667   .0362049     3.66   0.000     .0614919    .2034415
       _cons |   1.656612   .0670205    24.72   0.000     1.525227    1.787996
-------------+----------------------------------------------------------------
     sigma_u |  .40495036
     sigma_e |  .62554574
         rho |  .29531259   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(59, 5931) =    41.90                 Prob > F = 0.0000

. est store fe

. xtreg log_dt complex2 complex3 tau_d tau logev cti cti2 cti3 obj_diff, re

Random-effects GLS regression                   Number of obs      =      6000
Group variable: i                               Number of groups   =        60

R-sq:  within  = 0.2004                         Obs per group: min =       100
       between = 0.0002                                        avg =     100.0
       overall = 0.1491                                        max =       100

                                                Wald chi2(9)       =   1486.20
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      log_dt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    complex2 |   .2374744   .0416953     5.70   0.000     .1557531    .3191957
    complex3 |   .3771036   .0552865     6.82   0.000      .268744    .4854631
      tau_d  |  -.0022045   .0006464    -3.41   0.001    -.0034713   -.0009376
         tau |  -.0032846    .000323   -10.17   0.000    -.0039176   -.0026516
       logev |   .0601223   .0474916     1.27   0.206    -.0329595    .1532041
         cti |  -5.670031   .3304062   -17.16   0.000    -6.317615   -5.022446
        cti2 |   12.45948   1.223303    10.19   0.000     10.06185    14.85711
        cti3 |  -8.643612   1.142337    -7.57   0.000    -10.88255   -6.404672
    obj_diff |   .1324644   .0361992     3.66   0.000     .0615152    .2034136
       _cons |   1.656426   .0851303    19.46   0.000     1.489574    1.823278
-------------+----------------------------------------------------------------
     sigma_u |  .40678176
     sigma_e |  .62554574
         rho |  .2971941   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. est store re

. hausman fe re

                 ---- Coefficients ----
                 |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
                 |       fe           re         Difference          S.E.
-----------------+------------------------------------------------------------
    complex2     |   .2374196    .2374744       -.0000548        .0007429
    complex3     |   .3770243    .3771036       -.0000792        .0009906
      tau_d      |  -.0022045   -.0022045       -3.38e-08        .0000115
         tau     |  -.0032846   -.0032846        8.98e-17        5.72e-06
       logev     |   .0601018    .0601223       -.0000205        .0008433
         cti     |  -5.671949   -5.670031       -.0019184        .0082029
        cti2     |   12.46305    12.45948          .00357        .0274092
        cti3     |  -8.645294   -8.643612       -.0016819        .0239862
    obj_diff     |   .1324667    .1324644        2.26e-06        .0006416
------------------------------------------------------------------------------
```

is

each subject. It simply recog
te only their *variance*, $\sigma_u^2$. Note
cts, because there are far fewer
ndom effects if this model turns

anel data model required at all
led regression model estimated
of testing for differences within
fects in the fixed-effects model

$= 0$

on the model. The test is rou
e fixed-effects model. If this null
nclude that there are significant
ed regression model invalid and

dom effects (RE), the Hausman
ne two sets of estimates. The test
roughly expressed the reasoning
more stringent than those under
each other, this implies that the
alid, and RE is preferred because
stimates are very different from
, and the assumptions underlying
ts in a rejection, FE is preferred

nmands in STATA always start
) regression is carried out using
ators are implemented using the

usman test performed, with the

```
 cti cti2 cti3 obj_dif. fe

Number of obs      =
Number of groups   =

Obs per group: min =
               avg =
               max =

                   =
F(9,5931)          =
Prob > F           =
```

```
                    b = consistent under Ho and Ha; obtained from xtreg
     B = inconsistent under Ha, efficient under Ho; obtained from xtreg

  Test:  Ho:  difference in coefficients not systematic

                chi2(9) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                        =          0.17
              Prob>chi2 =        1.0000
```

We see that the two sets of results are indeed very similar, and it is not surprising that the Hausman test leads us to favour the random effects model (with a p-value of 1.0000). In the following section, we shall therefore focus on the results of the random effects model in interpretation.

## 5.6 Discussion of Results

For ease of comparison, we present the results from all estimated models in Table 5.2.

Since the F-test strongly rejects the pooled OLS model ($p = 0.0000$), and the Hausman test indicates acceptance of the random effects model ($p = 1.0000$), we shall interpret the results from the random effects model, shown in the final column of Table 5.2.

| Log decision time | OLS | Cluster OLS | Fixed effects | Random effects |
|---|---|---|---|---|
| Constant | 1.649(0.073) | 1.649(0.071) | 1.656(0.067) | 1.656(0.085) |
| Complexity 1 (base) | — | — | — | — |
| Complexity 2 | 0.240(0.049) | 0.240(0.040) | 0.237(0.042) | 0.237(0.042) |
| Complexity 3 | 0.380(0.065) | 0.380(0.053) | 0.377(0.055) | 0.377(0.055) |
| $\tau^d$ | −0.002(0.0008) | −0.002(0.0008) | −0.002(0.0006) | −0.002(0.0006) |
| $\tau$ | −0.003(0.0004) | −0.003(0.0004) | −0.003(0.0003) | −0.003(0.0003) |
| $Log(EV)$ | 0.061(0.056) | 0.061(0.048) | 0.060(0.047) | 0.060(0.047) |
| $cti$ | −5.591(0.389) | −5.591(0.357) | −5.672(0.331) | −5.670(0.330) |
| $cti^2$ | 12.313(1.443) | 12.313(1.283) | 12.463(1.223) | 12.459(1.223) |
| $cti^3$ | −8.576(1.350) | −8.576(1.178) | −8.645(1.143) | −8.643(1.142) |
| $\Delta^o$ | 0.132(0.043) | 0.132(0.034) | 0.132(0.036) | 0.132(0.036) |
| $\sigma_\epsilon$ | 0.741 | 0.741 | 0.625 | 0.625 |
| $\sigma_u$ | — | — | 0.405 | 0.407 |
| $n$ | 60 | 60 | 60 | 60 |
| $T$ | 100 | 100 | 100 | 100 |
| F-test (59, 5931) | | 41.90 ($p$=0.0000) | | |
| Hausman test $\chi^2(9)$ | | | | 0.17($p$=1.0000) |

Table 5.2: Results from log-linear regression models for decision time
Note: Standard errors in parentheses.

Firstly we note from the estimate of the intercept that the predicted decision time for the "easiest" type of problem, that is, a problem of complexity Level 1, for which the two lotteries are in fact identical ($\Delta^o = 0$), is exp(1.656) = 5.238 seconds. Of course, this problem must also be assumed to be right at the start of the experiment ($\tau^d = \tau = 0$) and this might explain why the prediction appears so high for such a simple problem.

Secondly we note that all included explanatory variables, except one, show strong significance. The effect of complexity is as expected: problems involving more outcomes take longer to solve. We do notice that the extra effort induced in moving from Level 2 to Level 3 is lower than that induced in the move from 1 to 2. This might be interpreted as evidence that subjects are discouraged by complex tasks, and this interpretation could be extended to predict a reduction of effort when complexity reaches intolerable levels. However, there is no evidence that such levels of complexity are being encountered in this experiment.

The effect of experience is quite dramatic. Both $\tau$ and $\tau^d$ are seen to have a strongly significant negative effect on decision time. As previously suggested, the first of these is interpretable as an experience effect, and the second as a boredom effect.

The coefficient of the variable $log(EV)$, 0.060, may be interpreted as an elasticity of effort with respect to financial incentives: if all prizes were doubled, we would expect response times (effort) to rise by around 6%. This is consistent with the view of many economists concerning incentives: that higher incentives bring about an increase in effort expended. However, we see that this effect is not actually significant, so we do not have statistical confirmation of this prior belief.

Finally, and perhaps most importantly, we see a strong negative and non-linear effect of closeness to indifference coupled with a strong positive effect of objective difference. This is exactly what we expected on the basis of our theoretical model of effort allocation described in Section 5.3. Figure 5.7 demonstrates the first effect clearly. Here, we have used the model estimates to predict the decision time at each value of closeness-to-indifference, for different complexity levels, and with other explanatory variables set to representative values. We see from Figure 5.7 that as the subject approaches indifference, the decision time rises steeply, and when a subject is actually indifferent (i.e. when closeness-to-indifference = 0), the predicted response time is more than double what it would be if one alternative were clearly preferred.

Of course, the pattern seen in Figure 5.7 is not surprising given the very similar shape seen in the non-parametric regression in Figure 5.6. This similarity is consistent with the correctness of the specification of our effort model (5.2)

## 5.7 Post-estimation

In Section 5.6 we established that the random effects model is the most suitable model for analysing the data set on decision times, and we interpreted the results from this model. The random effects approach is a natural means of capturing

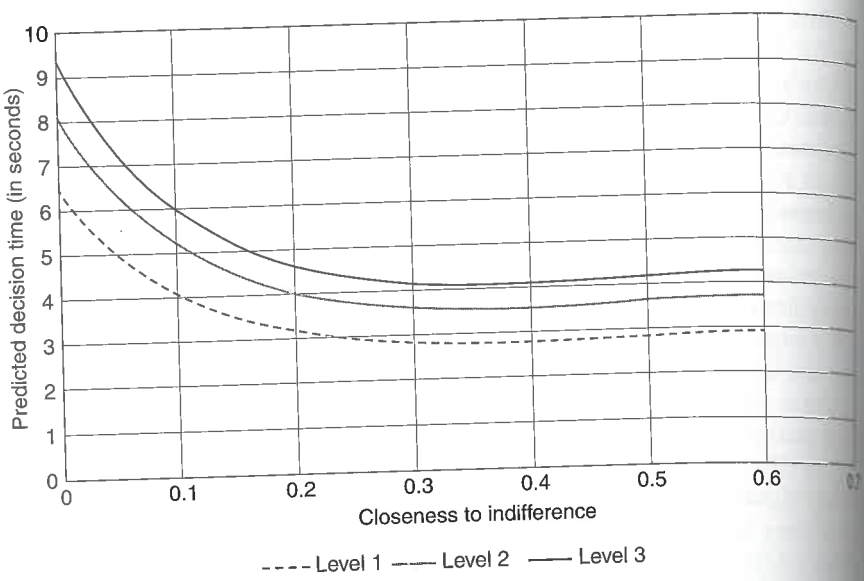



Figure 5.7: Predicted decision time against absolute valuation differential, at different complexity levels

Notes: $\tau$ set to 0; EV set to 10; $\Delta^0$ set to 0.5.

between-subject heterogeneity, and it is an approach that is followed repeatedly throughout the book. Sometimes, the random effect term has a clear interpretation. For example, in the risky choice model estimated in Chapter 13, the random effect term may be interpreted as the subject's coefficient of risk aversion.

Having estimated a random effects model, a natural question to ask is: what is the (estimated) value of the random effect for each subject? We shall refer to these values as the posterior random effects. This is a task that is carried out routinely following estimation. It usually involves the application of Bayes' rule to the estimated parameters together with the data. After using the `xtreg` command, only one additional command is required to generate the posterior random effects. The command is:

```
predict u_hat, u
```

This command stores the posterior random effects in the new variable u_hat. We present descriptive statistics of this variable below, and a histogram of u_hat is shown in Figure 5.8. Note that, for these purposes, we only require one observation for each subject; hence, the use of `if t==1` in the command.

```
. summ u_hat if t==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       u_hat |        60    5.59e-10    .3955946  -1.093639   .7443957
```

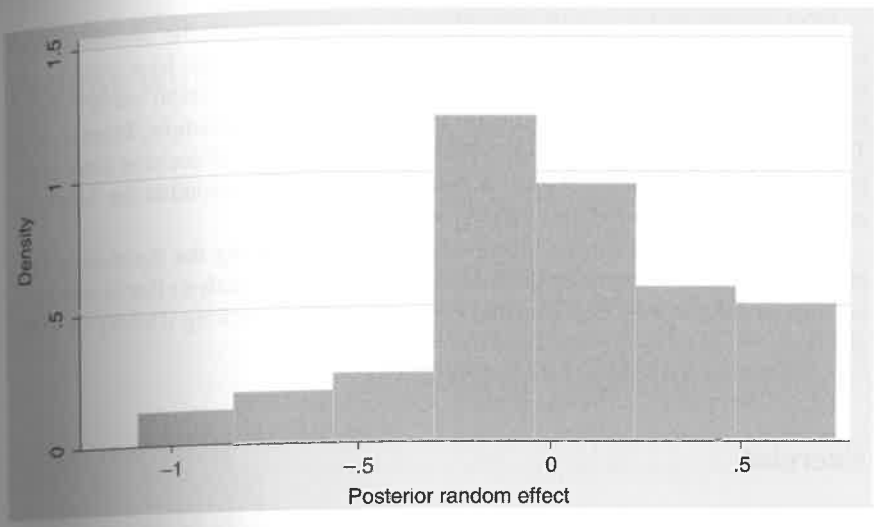



Figure 5.8: A histogram of posterior random effects

As expected, the posterior random effect has a mean of zero, and a standard deviation very close to the estimate of $\sigma_u^2$ reported in the random effects results, which was 0.407.

In the discussion of Figure 5.4, it was noted that subject 24 is the fastest decision maker, with a mean decision time of 1.558 seconds. It is therefore not surprising that this subject has the lowest posterior random effect, of −1.09. It was also noted that subject 29 is the slowest decision maker, with a mean decision time of 9.947 seconds. Again, it is not surprising that this subject has the highest posterior random effect, of +0.744.

## 5.8 Summary and Further Reading

This chapter has applied regression analysis to the modelling of decision times, with a view to identifying the determinants of effort in experimental tasks. Similar goals have been pursued by Wilcox (1994), who finds evidence that the use of simple rules in valuation tasks is associated with a reduction in decision time, suggesting that the use of such rules are motivated by a desire to save effort. Hey (1995) identifies factors which influence decision time in risky choice tasks, and then estimates choice models which allow the "noisiness" of response to depend on these factors. Decision times have also been analysed by Buschena & Zilberman (2000), Moffatt (2005*b*), and Alos-Ferrer et al. (2012).

The data used in this chapter was simulated. However, the simulation was created to resemble the real data set of Hey (2001) analysed by Moffatt (2005*b*). Readers interested to know how the simulation was performed are referred to Sections 13.3.2 and 13.4.6 of this book.

This seems to be a good example of an area in which the two disciplines of experimental economics and psychology overlap. Economists have traditionally been interested in the decisions made, not in the process by which the decision is reached. Psychologists are interested in the process (see, for example, Busemeyer & Townsend, 1993). The recent surge of interest in the analysis of decision times in the experimental economics literature is perhaps a sign that economists are becoming more interested in the decision-making process.

Of course there are more informative ways of analysing the decision-making process than simply observing decision times. One type of analysis that is starting to become popular among experimental economisis is eye-tracking (Holmqvist et al., 2011).

## Exercise

Little (1949, p. 92) posed the following question: "how long must a person dither before he is pronounced indifferent?" How can the results presented in Section 5.6, and particularly Figure 5.7, be invoked to provide an answer to this question?