EXPERIMETRICS

Peter G MOFFATT

National Taiwan University

April 2019

Contents

1 Power Analysis

1.1 Revision of Hypothesis Testing

The theme of the first part of this course is the use Power Analysis in Experimental Economics. There is no doubt that power analysis is now being taken very seriously by the field journals. For example, in the Editor's Preface of the very first issue (July, 2015) of the *Journal of the Economic Science Association*, the message is very clear: "A necessary (but not sufficient) condition for publishing a replication study or null result will be the presentation of power calculations."[1]

This section of the course provides a brief review of the central concepts underlying treatment testing, that are necessary for an understanding of power analysis. For further detail on these concepts see Siegel & Castellan (1988).

1.1.1 Key concepts and definitions

- A treatment test always has a *null hypothesis*, labelled H_0, and an *alternative hypothesis*, labelled H_1. The null hypothesis is typically the hypothesis that there is no effect. The alternative hypothesis is that there is an effect.

- The *true* size of the effect is referred to as the *effect size*. For example, if the value of the parameter of interest is θ_0 under the null, and θ_1 under the alternative, then the *effect size* is $\theta_1 - \theta_0$.

- 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 n data values in the sample. n is the *sample size*.

- The second stage is to compare the test statistic to the *null distribution* (i.e. the distribution that the statistic would in theory follow if the null hypothesis were true). The tails of this distribution form the *rejection region* of the test, and if the test statistic falls in this region, the null hypothesis is rejected in favour of the alternative. If the test statistic falls elsewhere, the null hypothesis is not rejected, and it may be concluded that the test result is consistent with the null hypothesis.

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

- The *p-value* of the test is the probability of obtaining a test statistic that is at least as extreme than the one obtained. One reason why the p-value is useful

[1] https://link.springer.com/article/10.1007/s40881-015-0012-4

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 main reason why the p-value is useful is because it is a measure of the *strength of evidence* against the null, and in favour of the alternative (i.e. 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.

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

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

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

- The *power* of a test is the probability of rejecting the null hypothesis when it is false. The power is denoted as π. Note that $\pi = 1 - \beta$.

- The power of a test is determined by a number of factors, including the *true effect size*, the sample size (n), and whether the test is one-tailed or two-tailed. It also depends on the chosen value of α. The higher α is, the higher the probability of type 1 error, which has the benefit of higher power.

- *Power analysis* is the name given to the set of techniques used to compute the power of a given test, and to find the sample size required to meet a given power requirement.

1.1.2 Choosing the value of α

The second last bullet point above tells us that the choice of α is an important decision. This choice depends to a large extent on the type of hypothesis under test. For example, first consider a situation in which the null hypothesis is that a crime suspect is not-guilty, and the alternative is that the suspect is guilty. In this situation, a type 1 error is finding an innocent person guilty, while a type 2 error is letting a guilty person go free. Many people view the first error as more serious than the second. Hence we should choose a very low value of α in this situation. How low? This is another question, although please note that α cannot be lowered all the way to zero, since this would mean that every suspect must be declared not-guilty.

Secondly, consider a situation in which the null hypothesis is that a patient is healthy, and the alternative is that they are suffering from a contagious disease. In this situation, declaring a diseased patient healthy (type 2 error) might be considered much more serious than telling a healthy patient that they have the disease (type 1 error). Hence, in this situation we could allow a higher value of α, since this would give rise

to a higher probability of detecting infected patients (i.e. higher power).

In Experimental Economics, we are perhaps fortunate that the errors that might be made in the interpretation of results from hypothesis tests rarely have consequences that are very serious. Hence it seems reasonable to follow convention and set the value of α to 0.05 as a standard.

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 β is 0.2. These conventions imply a four-to-one ratio between the probability of type II error and the probability of type I error. So, back to the example of the suspect in court, the number of guilty suspects set free is four times the number of innocent suspects imprisoned.

1.2 Treatment Testing

We will demonstrate a number of standard treatment testing techniques in a particular context: the "Willingness to Pay – Willingness to Accept Gap", or just the "WTP – WTA Gap". This phenomenon has been studied extensively using experimental data. See, for example, Kahneman et al. (1990), Plott & Zeiler (2007) and Isoni et al. (2011). We will use examples taken directly from Isoni et al. (2011). The application is particularly useful for the demonstration of treatment tests because a range of different types of test are required, and can therefore be demonstrated naturally.

Experiments on the WTP-WTA Gap sometimes require subjects to value physical objects, and sometimes lotteries. We will consider both of these, commencing with the former.

1.2.1 One–sample Tests

Before considering treatment tests, we shall consider one–sample tests, since these are the tests we start with when considering power calculations later.

The experiment of Isoni et al. (2011) has 100 subjects. One of the tasks requires subjects to value a coffee mug, whose retail value is £3.00.[2] An obvious initial question to address is how close the valuations are to the market price; if valuations are close to the market price, this indicates that the market price is an accuarate reflection of individuals' valuations. To address this question, it is appropriate to set up a hypothesis test of $H_0 : \mu = 3.00$, against $H_1 : \mu \neq 3.00$. It is important to recognise that we are using the *sample* of 100 valuations to test the hypothesis that the *population mean valuation* is equal to 3.00.

The test that is required here is the one-sample t-test. The test statistic is:

[2] If you read Isoni et al. (2011), you will find that the retail value of the mug is in fact £4.50. We are pretending that the price is £3.00 in order to make the results of the one-sample test easier to interpret.

$$t = \frac{\bar{y} - 3.00}{s/\sqrt{n}} \tag{1}$$

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 (5) has a $t(n-1)$ distribution. Hence, the rejection rule, given our chosen value of α, is $|t| > t_{n-1,\alpha/2}$.

Summary statistics for the valuations are obtained as follows: do-file_1.do

```
. summ v_mug

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       v_mug |        100      2.0415    1.571287          0        7.5
```

Inserting the summary statistics into the the formula for the t-test (1), we obtain the test statistic:

$$t = \frac{2.0415 - 3.00}{1.5713/\sqrt{100}} = -6.10 \tag{2}$$

We then compare this test statistic to the $t(99)$ distribution. Since it is a 2-tailed test, we use the critical value $t_{99,0.025} = 1.99$. We reject H_0 in favour of H_1 because $|-6.10| > 1.99$. We conclude that we have evidence that the (population) mean valuation is different from the market price of 3.00.

Note that this test can be performed in STATA using the `ttest` command:

```
. ttest v_mug=3.0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   v_mug |     100      2.0415    .1571287    1.571287    1.729723    2.353277
------------------------------------------------------------------------------
    mean = mean(v_mug)                                            t =  -6.1001
Ho: mean = 3.0                                 degrees of freedom =       99

   Ha: mean < 3.0               Ha: mean != 3.0                 Ha: mean > 3.0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Of course this gives the same results. Note that in addition STATA gives three p-values. The one we are interested in is the one in the centre, which corresponds to the 2-tailed test. This p-value of 0.0000 tells us that the evidence that the population mean valuation is different from 3 is overwhelming.

1.2.2 Between-subject (parametric) Treatment Tests

Isoni et al. (2011)'s 100 subjects are randomly allocated between a WTA (51) and WTP(49) treatments.[3] The sample means of WTA and WTP are £2.21 and £1.86 respectively. In investigating the WTP – WTA Gap, the key question to be asked is whether this difference is statistically significant.

It is useful to start by plotting the two distributions in histograms. This is done in Figure 1. What is perhaps the most striking feature of these graphs is the difference

[3]WTP and WTA are elicited using the Becker-DeGroot-Marschak incentive mechanism. See Becker et al. (1964).

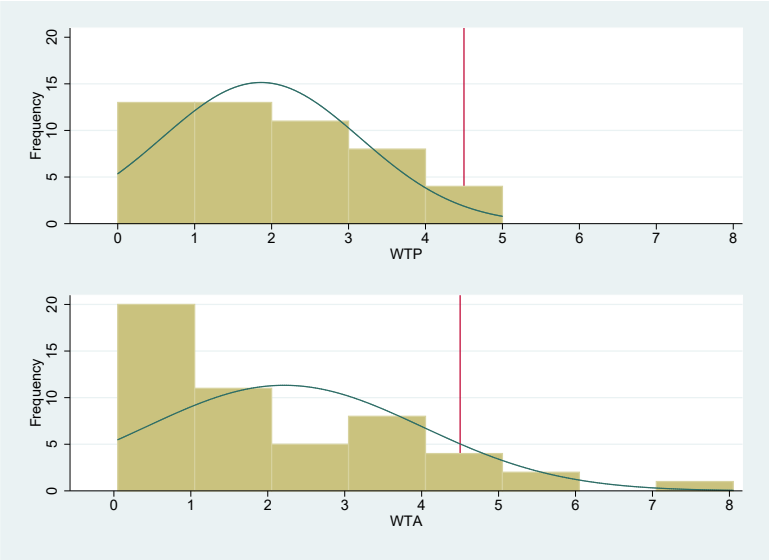


Figure 1: Frequency histograms of WTP (n=49) and WTA (n=51) for a coffee cup. Data from Isoni et al. (2011). Vertical line at retail value (£4.50). Normal densities superimposed.

Peter conjectures that the spread difference comes from people being more familiar with buying a mug, compared to selling a mug. This actually predicts that you could reverse the WTP, WTA spread difference by framing it as buying or selling labor.

in spead: the spread of values is clearly higher for WTA than for WTP. It is a simple matter to test the significance of this difference, by using the variance ratio test, performed in STATA with the `sdtest` command. The results are as follows:

```
. sdtest v_mug, by(v_type)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTP |      49    1.862245     .184379    1.290653    1.491526    2.232964
     WTA |      51    2.213725    .2515679    1.796554    1.708436    2.719015
---------+--------------------------------------------------------------------
combined |     100      2.0415    .1571287    1.571287    1.729723    2.353277
------------------------------------------------------------------------------
    ratio = sd(WTP) / sd(WTA)                                     f =   0.5161
Ho: ratio = 1                                 degrees of freedom =   48, 50

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0114         2*Pr(F < f) = 0.0229          Pr(F > f) = 0.9886
```

The test statistic for this test is simply the ratio of the two variances, or equivalently the square of the ratio of the two standard deviations. To be explicit, the test statistic, F, has been computed as:

$$F = \frac{s_1^2}{s_2^2} = \frac{1.29^2}{1.79^2} = \underline{0.52} \tag{3}$$

This statistic follows an $F(48, 50)$ distribution under the null hypothesis that the two variances are equal. In this case the test statistic is 0.52. Based on a 2-tailed test, we see that there is evidence of a difference between the two variances (p-value=0.0229).

We next perform an independent samples t-test of the null hypothesis that the two means are equal. However, since we have just discovered that the variances differ between the two samples, it is important to perform the version of the t-test that allows the variances to be unequal. The formula for the test statistic is:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the two sample means, s_1^2 and s_2^2 are the two sample variances, and n_1 and n_2 are the numbers of observations in each treatment.

This test is performed in STATA with the `ttest` command. The `unequal` option is required because the variances must be assumed to be unequal. The results are:

```
. ttest v_mug, by(v_type) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTP |      49    1.862245     .184379    1.290653    1.491526    2.232964
     WTA |      51    2.213725    .2515679    1.796554    1.708436    2.719015
---------+--------------------------------------------------------------------
combined |     100      2.0415    .1571287    1.571287    1.729723    2.353277
---------+--------------------------------------------------------------------
    diff |           -.3514806    .3119007               -.9710476    .2680864
------------------------------------------------------------------------------
    diff = mean(WTP) - mean(WTA)                                  t =  -1.1269
Ho: diff = 0                     Satterthwaite's degrees of freedom =  90.8403

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1314         Pr(|T| > |t|) = 0.2628          Pr(T > t) = 0.8686
```

Since we have a prior belief that $WTP < WTA$, we may perform a one-tailed ($<$) test. This means that we take the p-value of 0.1314. However, this does not represent evidence of a difference between the two means. It appears that this particular experiment does not provide evidence of a $WTP - WTA$ Gap, on the basis of the parametric t-test.

There are a number of other issues to address. Referring back to Figure 1, we see that neither of the two distributions appears close to the superimposed normal densities. Non-normality of the data can be confirmed using various statistical tests. One simple test is the skewness-kurtosis test performed in STATA with the sktest command. The results are:

\texttt{sktest}

```
. sktest v_mug if v_type==0

                    Skewness/Kurtosis tests for Normality
                                                      ------ joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
-------------+-------------------------------------------------------------
       v_mug |         49     0.1143        0.1334        4.74         0.0936

. sktest v_mug if v_type==1

                    Skewness/Kurtosis tests for Normality
                                                      ------ joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
-------------+-------------------------------------------------------------
       v_mug |         51     0.0133        0.6210        5.97         0.0504
```

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[4] equals zero

[4] *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.

(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".[5]

One clear departure from normality is the skewness in WTA (p-value=0.013). The two overall tests give us only mild evidence of non-normality.

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

1.2.3 Treatment testing using a regression

Consider the independent samples t-test performed in the last Section, but this time disregard the complication of unequal variances.

```
. ttest v_mug, by(v_type)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTP |      49    1.862245     .184379    1.290653    1.491526    2.232964
     WTA |      51    2.213725    .2515679    1.796554    1.708436    2.719015
---------+--------------------------------------------------------------------
combined |     100      2.0415    .1571287    1.571287    1.729723    2.353277
---------+--------------------------------------------------------------------
    diff |           -.3514806    .3139184               -.9744414    .2714802
------------------------------------------------------------------------------
    diff = mean(WTP) - mean(WTA)                                  t =  -1.1197
Ho: diff = 0                                     degrees of freedom =       98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1328         Pr(|T| > |t|) = 0.2656          Pr(T > t) = 0.8672
```

An important point is that this t-test can be performed using a linear regression, with valuation as the dependent variable, and a dummy variable representing the treatment as the single explanatory variable:

```
. regress v_mug v_type

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =      1.25
       Model |  3.08722984         1  3.08722984   Prob > F        =    0.2656
    Residual |  241.338048        98  2.46263314   R-squared       =    0.0126
-------------+----------------------------------   Adj R-squared   =    0.0026
       Total |  244.425278        99   2.4689422   Root MSE        =    1.5693
```

[5] *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*.

```
------------------------------------------------------------------------------
       v_mug |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      v_type |   .3514806   .3139184     1.12   0.266    -.2714802     .9744414
       _cons |   1.862245   .2241826     8.31   0.000     1.417362     2.307128
------------------------------------------------------------------------------
```

Note that the t-statistic associated with the treatment dummy is the same (in magnitude) as the t-statistic obtained using the `ttest`. It is important to recognise this equivalence when performing treatment tests in the context of a regression model.

1.2.4 Between-subject (non-parametric) Treatment Tests

Suppose that we have reason to assume that the t-test considered in the last section is unreliable. The natural alternative is a non-parametric test. Perhaps the most popular non-parametric treatment test among Experimental Economists is the Mann-Whitney test. It is classified as non-parametric because it does not rely on any strong distributional assumptions (such as normality of the data).

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

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

The test is carried out in STATA using the `ranksum` command. The result from applying the test to the $WTP - WTA$ data is:

```
. ranksum v_mug, by(v_type)

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

      v_type |      obs    rank sum    expected
-------------+---------------------------------
         WTP |       49      2392.5      2474.5
         WTA |       51      2657.5      2575.5
-------------+---------------------------------
    combined |      100        5050        5050

unadjusted variance    21033.25
adjustment for ties      -90.75
                     ----------
adjusted variance      20942.50

Ho: v_mug(v_type==WTP) = v_mug(v_type==WTA)
             z =  -0.567
    Prob > |z| =   0.5710
```

I believe the ranksum command in STATA uses the normal approximation when calculating p-values. Is there an exact option like MATLAB?

Try "porder"

The p-value of 0.5710 indicates that there is no evidence of a $WTP - WTA$ Gap. We also see from the p-value that the evidence of a Gap is even weaker than that from the t-test in tha last section. This is actually an expected result: evidence of an effect tends to be weaker, the less is assumed about the process generating the data.

1.2.5 Tests comparing entire distributions

As suggested by Forsythe et al. (1994), tests comparing entire distributions are useful in situations in which economic theory does not predict the precise nature of the treatment effect. In the context of the current example, since elementary consumer theory predicts the equality of WTP and WTA, the same theory is not useful in predicting the nature of any deviation of WTA from WTP. More precisely, theory does not predict which functional of the distribution may be expected to shift in response to the WTP/WTA treatment. Is it (as usually assumed) the mean of the distibution that shifts? Or is it the median? Or is it the spread of the distribution (and Figure 1 provided evidence that it might well be this)? This problem is solved by applying tests that are based on a comparison of the entire distributions under the two treatments, rather than a comparison of a particular functional such as mean or variance.

One popular test that compares two entire distributions is the Kolmogorov-Smirnov (KS) test. In order to understand this test, it is useful to present the two cumulative distribution functions (cdf's) on the same graph. Such a graph is shown in Figure 2. The observation that the WTA cdf lies broadly below and to the right of the WTP cdf is consistent with WTA being higher than WTP. The KS test statistic is used to judge whether this difference is significant. With reference to Figure 2, the KS test statistic is seen to be something very simple: it is the maximum vertical distance between the two cdf's. This is in fact +0.1309, which is the vertical distance between the two cdf's when the value of mug is between 3.0 and 3.4.

"cdfplot" is a user-written command; use "find it cdfplot" to find and install that on your STATA.

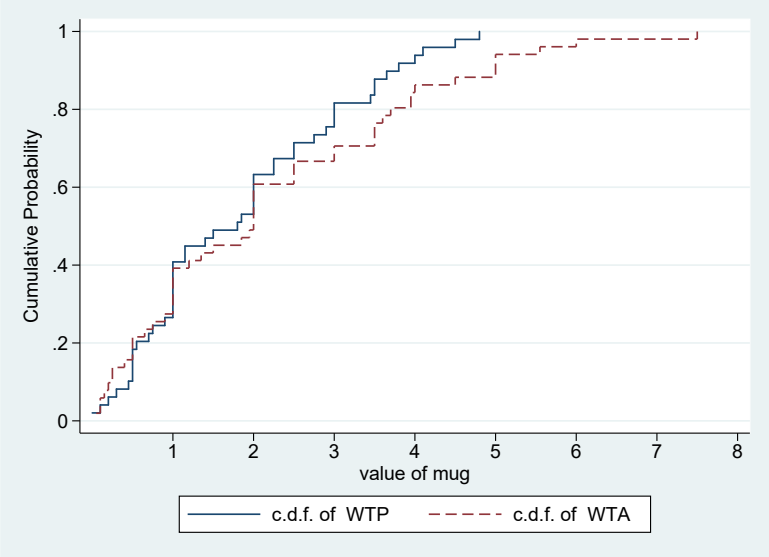


Figure 2: The cdf's of WTP and WTA.

The KS test is implemented by the command `ksmirnov` in STATA. Applied to the WTP-WTA data, the results are as follows:

```
. ksmirnov v_mug, by(v_type)

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

 Smaller group       D       P-value
```

```
-----------------------------------
WTP:                0.1309     0.425
WTA:               -0.0760     0.749
Combined K-S:       0.1309     0.786

Note: Ties exist in combined dataset;
      there are 45 unique values out of 100 observations.
```

There are three test statistics (D), and accompanying p-values. The first is the maximum vertical distance of the WTP cdf over the WTA cdf, as already discussed. The second is the maximum distance of WTA over WTP which is much smaller. The third is a combined test-statistic which may be used for a 2-sided test. We see that whichever of the three tests is used, there is no evidence of a difference between the two distributions, and hence no evidence of a WTP-WTA gap.

Discrete K-S test

Another test for comparing entire distributions that has become quite popular in Experimental Economics is the Epps-Singleton test (Epps and Singleton, 1986). In fact, this test does not compare the two distributions directly, but instead compares the empirical characteristic functions. 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).

$\phi(t) = E\left(e^{itX}\right)$

```
escftest v_mug, group(v_type)
```

1.2.6 Within-subject tests

Until now, it has been assumed that the two treatments in a treatment test have been administered to two samples separately. Such tests are known as between-subject tests. Within-subject tests are used to test the effect of a treatment in the contrasting situation in which each subject is observed both before and after the treatment.

From a theoretical point of view, within-subject tests are preferred to between-subject tests, for the obvious reason that they have greater statistical power. 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 two treatments 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. For example, while in a WTP-WTA comparison in the context of a physical object (such as coffee mugs) there are practical reasons for not exposing both treatments to a single subject, in the context of a lotteries involving money amounts, this is a natural approach to take.

Isoni et al. (2011) ask their 100 subjects to state their WTA for the lottery $(\$4, 0.3; \$0, 0.7)$, and then in a later task, ask the same subjects to state their WTP for the lottery $(\$5, 0.3; \$1, 0.7)$. Note that these two lotteries are not identical, but the only difference is that the money amounts in the WTP lottery are exactly £1 more than the corresponding amounts in the WTA lottery. Hence the $WTP - WTA$

gap can reasonably be measured by comparing $WTA + \$1$ with WTP. For convenience we will henceforth refer to $WTA + \$1$ as WTA.

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 between WTA and WTP for each subject, and then applies the t-test to test whether these differences have mean zero. The results are:

```
. ttest WTA=WTP

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTA |     100      2.5745    .0961546    .9615459    2.383708    2.765292
     WTP |     100      2.2415    .1115095    1.115095    2.020241    2.462759
---------+--------------------------------------------------------------------
    diff |     100        .333    .1398705    1.398705    .0554666    .6105334
------------------------------------------------------------------------------
     mean(diff) = mean(WTA - WTP)                                      t =   2.3808
 Ho: mean(diff) = 0                               degrees of freedom =       99

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9904         Pr(|T| > |t|) = 0.0192          Pr(T > t) = 0.0096
```

We see that there is strong evidence (p-value=0.0096) of WTA being greater than WTP.

The non-parametric test appropriate in this situation is the Wilcoxon signed ranks test (see Siegel and Castellan, 1988). As with the parametric test, this test is based on the differences between WTA and WTP 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 there is no WTP-WTA Gap, 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 WTA=WTP

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+--------------------------------
    positive |       58      3315.5        2492
    negative |       31      1668.5        2492
        zero |       11          66          66
-------------+--------------------------------
         all |      100        5050        5050

unadjusted variance    84587.50
adjustment for ties      -22.38
adjustment for zeros    -126.50
                     ----------
adjusted variance      84438.63

Ho: WTA = WTP
             z =   2.834
    Prob > |z| =   0.0046
```

The rank sum for the positive differences is clearly a higher number, at 3315.5. The test gives a (two-tailed) p-value of 0.0046 which represents strong evidence of

a $WTP - WTA$ Gap. The one-tailed p-value is 0.0023, providing strong evidence that WTA is greater. This p-value is, surprisingly, lower than 0.0096 obtained above from the corresponding parametric test, indicating that the evidence from the non-parametric test is stronger.

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 may be viewed as a fully non-parametric test. It too can be peformed in STATA:

```
. signtest WTA=WTP

Sign test

        sign |    observed    expected
-------------+------------------------
    positive |          58        44.5
    negative |          31        44.5
        zero |          11          11
-------------+------------------------
         all |         100         100

One-sided tests:
  Ho: median of WTA - WTP = 0 vs.
  Ha: median of WTA - WTP > 0
      Pr(#positive >= 58) =
         Binomial(n = 89, x >= 58, p = 0.5) =  0.0028

  Ho: median of WTA - WTP = 0 vs.
  Ha: median of WTA - WTP < 0
      Pr(#negative >= 31) =
         Binomial(n = 89, x >= 31, p = 0.5) =  0.9986

Two-sided test:
  Ho: median of WTA - WTP = 0 vs.
  Ha: median of WTA - WTP != 0
      Pr(#positive >= 58 or #negative >= 58) =
         min(1, 2*Binomial(n = 89, x >= 58, p = 0.5)) =  0.0055
```

The relevant p-value is the first one, 0.0028. Once again there is strong evidence that WTA is higher than WTP.

1.3 Power Analysis - Theory

Power analysis (Cohen 2013) is used to find the power of a test that has been performed, power being defined as the probability of detecting an effect given that the effect really exists. It can also be used to find the sample size required to perform a test with a given power.

1.3.1 Power analysis for one-sample tests

One-sample tests are rarely used in Experimental Economics, but they are simpler to analyse than the more useful independent-sample tests. This is why we commence with one-sample tests.

Suppose that we are interested in the continously distributed outcome measure Y whose population mean is μ. Suppose further that we are interested in testing the null hypothesis $\mu = \mu_0$ against the alternative hypothesis $\mu = \mu_1$, where $\mu_1 > \mu_0$.[6] We plan to collect a sample of size n for this purpose, and we need to decide what n should be. Recall that, before we do this, we need to set two quantities. The first is the test size, α, which is the probability of rejecting the null hypothesis when it is true (or the probability of type I error). The second is the probability of failing to reject the null hypothesis when it is false (or the probability of type II error). This second probability is conventionally labelled β. Note that the probability of rejecting the null hypothesis when it is false is $1 - \beta$ and this is the *power* of the test. We shall denote power by π.

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

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

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \tag{5}$$

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 (5) has a $t(n-1)$ distribution. Hence, the rejection rule, given our chosen value of α, 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 $t > z_\alpha$, where z_α is the upper α critical value of the standard normal. This simplifies the analysis considerably.

The power of the test is given by:

$$\begin{aligned} P(t > z_\alpha | \mu = \mu_1) &= P\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}} > z_\alpha \mid \mu = \mu_1\right) = P\left(\bar{y} > \mu_0 + z_\alpha\left(s/\sqrt{n}\right) \mid \mu = \mu_1\right) \\ &= P\left(\frac{\bar{y} - \mu_1}{s/\sqrt{n}} > \frac{\mu_0 + z_\alpha\left(s/\sqrt{n}\right) - \mu_1}{s/\sqrt{n}} \mid \mu = \mu_1\right) \\ &= \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha\left(s/\sqrt{n}\right)}{s/\sqrt{n}}\right) \end{aligned} \tag{6}$$

To see the formula (6) at work, suppose that we have a sample of size 30, and 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. As usual, we set $\alpha = 0.05$ so that $z_\alpha = 1.645$. Then we apply formula (6) to obtain:

[6] 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.

$$\pi = \Phi\left(\frac{12 - 10 - 1.645 \times 5/\sqrt{30}}{5/\sqrt{30}}\right) = \Phi(0.54584) = 0.71 \tag{7}$$

With a sample of 30, we see that the power of the test is 0.71. The question that follows naturally is: what would the sample size need to be for the power to reach the desired 0.80?

If the desired power of the test is $1 - \beta$, we have, from the last line of (6):

$$\frac{\mu_1 - \mu_0 - z_\alpha s/\sqrt{n}}{s/\sqrt{n}} = z_\beta \tag{8}$$

Rearranging (8) we obtain:

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

Recalling that our chosen values of α and β are 0.05 and 0.20 respectively, we have $z_\alpha = 1.645$ and $z_\beta = 0.842$. Hence we may write the formula for the required sample size as:

$$n = \frac{6.185 s^2}{(\mu_1 - \mu_0)^2} \tag{9}$$

Once again 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 (9) to obtain:

$$n = \frac{6.185 \times 5^2}{(12 - 10)^2} = 38.66 \tag{10}$$

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 `power` can be used to perform the calculations leading to both (7) and (10). To obtain the power when the sample size is 30, (7), we use the following syntax, and obtain the following result:

```
. power onemean 10 12 , sd(5) n(30)  oneside

Estimated power for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
            N =        30
        delta =    0.4000
           m0 =   10.0000
           ma =   12.0000
           sd =    5.0000

Estimated power:

        power =    0.6895
```

The main arguments are "onemean" which indicates that a one-sample test is required, and 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; "oneside" indicates that a one-sided test is required; "n(30)" indicates that the sample size is 30.

Note that the power computed by STATA of 0.6895 is slightly lower than the power obtained in (7), and this is a consequence of the latter using the normal as an approximation for the t-distribution. STATA correctly assumes the t-distribution, so this lower number is, stricty speaking, the correct power.

To obtain the sample size required to achieve a power of 0.80, (10), the required syntax, and results, are:

```
 . power onemean 10 12 , sd(5)  oneside p(0.8)

Performing iteration ...

Estimated sample size for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.4000
           m0 =   10.0000
           ma =   12.0000
           sd =    5.0000

Estimated sample size:

            N =        41
```

Note that the only change from the previous use of the `power` command is that the sample size option `n(30)` is replaced by the power option `p(0.80)`.

Note that the required sample size is 41, in close agreement with the calculation performed in (10) (which gave 39). Again, the reason why the result obtained by STATA is slightly larger than the one obtained above using a hand calculator is that the latter uses the normal as an approximation for the t-distribution. STATA correctly assumes the t-distribution and as a result the number appearing in the numerator of (9) is slightly larger.

A very useful feature of the `power` command is the `graph` option. This enables us to plot power functions.

The following command is used to plot power against sample size for a range of alternative hypotheses. The result is shown in Figure 3.

```
power onemean 10 (10.5(0.5)12.5), sd(5) n(20(10)200) oneside graph
```

The following command is used to plot the sample size required to attain a various different powers (including 0.80) against the mean under the alternative. The result is shown in Figure 4.

```
power onemean 10 (10.5(0.25)12.5), sd(5) p(0.6(0.1)0.9)  oneside graph
```

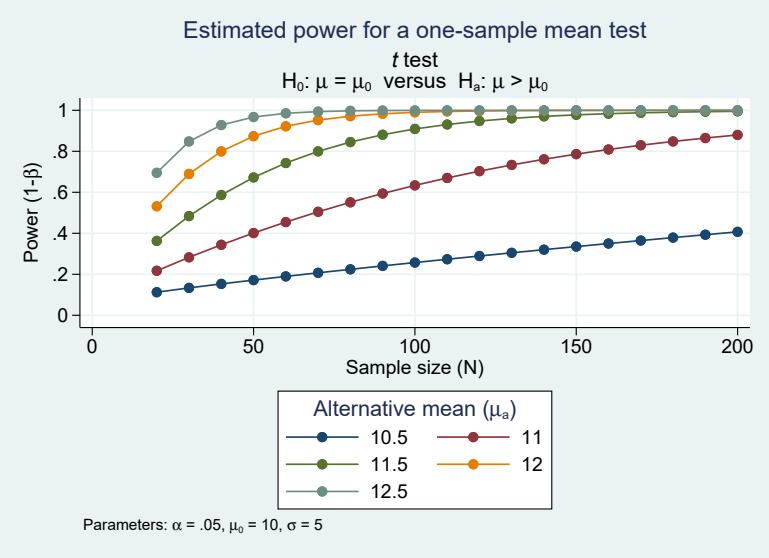


Figure 3: Power against sample size under different alternatives

1.3.2 Power analysis for two independent samples

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 μ_1 and μ_2 be the population means of the control group and the treatment group respectively. The null hypothesis of interest is $\mu_2 - \mu_1 = 0$ (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:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{11}$$

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

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{12}$$

The reason for using (12) is that we are assuming the two sub-samples have the same variance. A slightly different formula formula from (11) is required if the two variances are assumed to be unequal. Under the null hypothesis $\mu_2 - \mu_1 = 0$, the

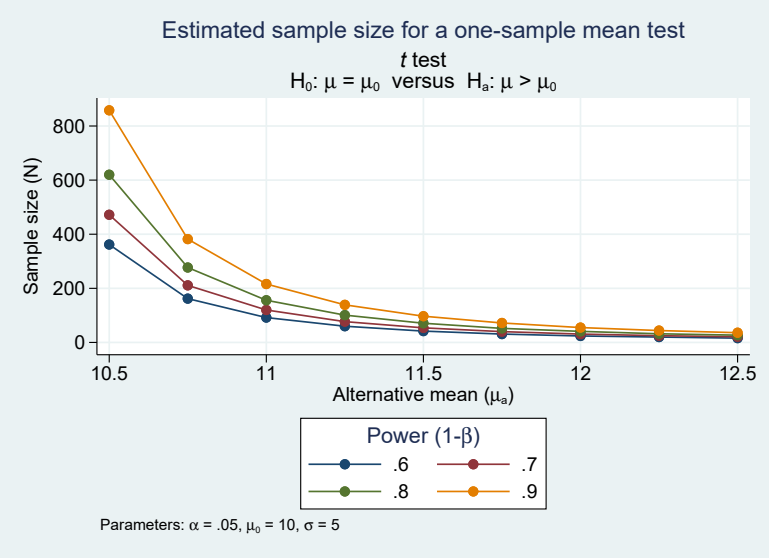


Figure 4: Required sample size (for various powers) against mean under the alternative

distribution of t given in (11) is $t(n_1 + n_2 - 2)$. Again, matters are simplified using the normal approximation. We will therefore use the critical value z_α.

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:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{2}{n}}} \tag{13}$$

The power of the test is given by:

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

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

$$\frac{d - z_\alpha s_p\sqrt{\frac{2}{n}}}{s_p\sqrt{\frac{2}{n}}} = z_\beta \tag{14}$$

Rearranging (14) we obtain:

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

Once again applying our chosen values of α and β, we have $z_\alpha = 1.645$ and $z_\beta = 0.842$, and we may write the formula for the required sample size as:

$$n = \frac{12.370 s_p^2}{d^2} \tag{15}$$

For an example of the use of formula (15), 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 (15) to obtain:

$$n = \frac{12.370 \times 25}{4} = 77.3$$

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

The STATA syntax for the test just performed is:

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

The main arguments are "twomeans" which indicates that a two-sample test is required, and the values of μ_1 and μ_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 close agreement with the calculation performed above.

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

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    2.0000
           m1 =   10.0000
           m2 =   12.0000
          sd1 =    4.0000
          sd2 =    5.8400

Estimated sample sizes:

            N =       158
  N per group =        79
```

Again the graph option is useful. The following command plots power against sample size for the test just performed. The result is shown in Figure 5. Note that the sample size measured on the x-axis of the graph is the total sample size (both groups). We see that that the total sample size giving a power of 0.8 is around 158, in agreement with the result of 79 per group obtained from the calculation.

```
power twomeans 10 12 , sd1(4.0) sd2(5.84) n(20(10)200) oneside  graph
```

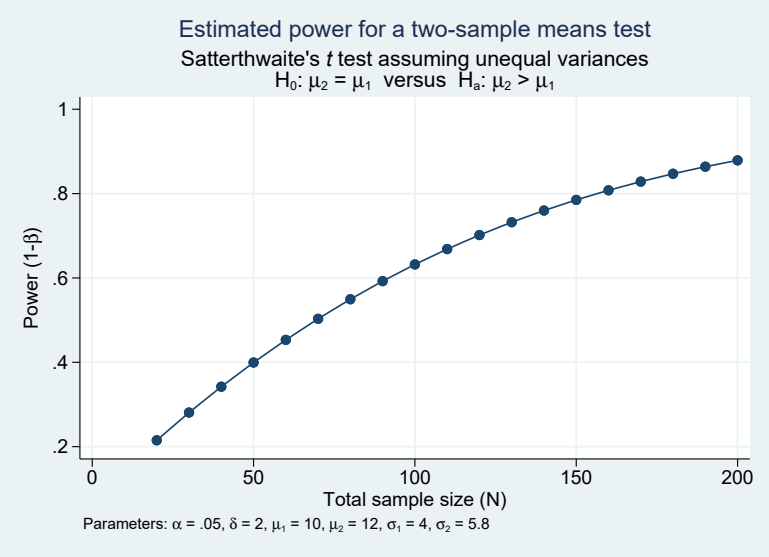


Figure 5: Power against sample size for treatment test

1.3.3 Power analysis for paired tests

In the last sub-section, we considered the problem of applying power analysis in the situation of an independent-samples t-test. Here, we consider the situation of a paired sample t-test. That is, if we assume that each subject is observed twice, once without the treatment, and once with the treatment, how do we find the number of subjects required to achieve a given power?

The theory used to derive the power formulae is very similar to that of Section 1.3.2, so there is no need to repeat it here. The syntax for the `power` command is slightly different.

Let us use the same example as used in Section 1.3.2: we are testing the effect size $d = 2$, and we know that the standard deviations of responses 1 and 2 are 4.0 and 5.84 respectively. The required command, and results, are as follows:

```
. power pairedmeans 10 12 , sd1(4.0) sd2(5.84) corr(0) oneside

Performing iteration ...

Estimated sample size for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d > d0

Study parameters:
```

```
        alpha =    0.0500          ma1 =   10.0000
        power =    0.8000          ma2 =   12.0000
        delta =    0.2825          sd1 =    4.0000
           d0 =    0.0000          sd2 =    5.8400
           da =    2.0000         corr =    0.0000
         sd_d =    7.0785

Estimated sample size:

            N =        79
```

We see that the required sample size is 79. Note that this is the same as the number of subjects required *in each group* in the independent samples test in Section 1.3.2. This is intuitive: asking 79 subjects to perform 2 tasks each is, in a sense, of equal value to asking 158 subjects to perform one task each.

However, note the use of the `corr(0)` option with the `power` command. This indicates that the two responses in the paired test are uncorrelated. It is likely that the two responses are positively correlated: If a subject's WTA is unusually high, it is reasonable to expect their WTP to be high as well.

Let us see what happens when we assume a positive correlation between the two responses:

```
. power pairedmeans 10 12 , sd1(4.0) sd2(5.84) corr(0.5) oneside

Performing iteration ...

Estimated sample size for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d > d0

Study parameters:

        alpha =    0.0500          ma1 =   10.0000
        power =    0.8000          ma2 =   12.0000
        delta =    0.3867          sd1 =    4.0000
           d0 =    0.0000          sd2 =    5.8400
           da =    2.0000         corr =    0.5000
         sd_d =    5.1716

Estimated sample size:

            N =        43
```

With a correlation of 0.5 assumed, the required sample size is considerably lower, at 43.

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

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

It should be added that a much-discussed disadvantage of paired designs is the possibility of "order effects", that is, the behaviour of subjects depending on the order in which the treatments are experienced. "Crossover designs" are a way of countering this problem: half of subjects see control followed by treatment; the other half see treatment followed by control. Differences between these two groups would confirm the existence of an order effect, which would then need to be controlled for in treatment tests.

1.4 Power analysis with real examples

1.4.1 Power analysis for the one-sample test on valuation data

In Section 1.2.1, we carried out a one-sample test of the hypothesis that the population mean valuation of a mug equals 3.0. Based on the available data, we found strong evidence that the mean is less than 3.0. Let us now apply power analysis to this testing problem, using the methods introduced in Section 1.3.1.

The first question is, what is the power of the test performed. Recall that the sample size is 100, and summary statistics for the 100 valuations are:

```
. summ v_mug

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       v_mug |        100      2.0415    1.571287          0        7.5
```

The result of the one-sample t-test of $H_0 : \mu = 3.0$ is:

```
. ttest v_mug=3.0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   v_mug |     100      2.0415    .1571287    1.571287    1.729723    2.353277
------------------------------------------------------------------------------
    mean = mean(v_mug)                                            t =  -6.1001
Ho: mean = 3.0                                 degrees of freedom =       99

   Ha: mean < 3.0               Ha: mean != 3.0                 Ha: mean > 3.0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

To find the power of this test, we apply the `power` command as follows:

```
. power onemean 3.0 2.04,n(100) sd(1.571)

Estimated power for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m != m0

Study parameters:

        alpha =    0.0500
            N =       100
        delta =   -0.6111
           m0 =    3.0000
           ma =    2.0400
           sd =    1.5710

Estimated power:

        power =    1.0000
```

We find that the power of this particular test is a very impressive 1.0! This means that with a sample of 100, and given that the true mean is 2.04 and the true standard deviation is 1.57, we will *always* find evidence that the mean is different from 3.0. This suggests that, if this were the hypothesis of interest, we could get away with a much smaller sample size. This leads us to the next question: what sample size would be required to obtain a power of 0.80. This requires a different `power` command:

```
. power onemean 3.0 2.04,p(0.80) sd(1.571)

Performing iteration ...

Estimated sample size for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m != m0

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   -0.6111
           m0 =    3.0000
           ma =    2.0400
           sd =    1.5710

Estimated sample size:

            N =        24
```

The result is that we would only need a sample of 24 to obtain a test with power 0.80.

1.4.2 Power analysis for the independent samples test on valuation data

To find the power of the independent samples t-test performed earlier on the WTP-WTA data, we use the `power` command in STATA, as follows:

```
. power twomeans 1.86 2.21 , n1(49) n2(51) sd1(1.29) sd2(1.80) oneside

Estimated power for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
            N =       100
           N1 =        49
           N2 =        51
        N2/N1 =    1.0408
        delta =    0.3500
           m1 =    1.8600
           m2 =    2.2100
          sd1 =    1.2900
          sd2 =    1.8000

Estimated power:

        power =    0.2973
```

The question that is being asked here is: if the true means of the two distributions were 1.86 and 2.21, and if the true standard deviations also happened to equal the sample standard deviations, and if we had group samples of size 49 and 51, then what would be the probability of rejecting the null hypothesis of no difference in

means, if we used the independent samples t-test? The computed power of 0.2973 seems low.

The other question to be addressed is: what sample sizes would be required to attain the benchmark power of 0.80? This essentially requires inversion of the formula that was used to compute power. The `power` command is used again, but with a different set of options:

```
. power twomeans 1.86 2.21 , sd1(1.29) sd2(1.80) oneside power(0.8)

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.3500
           m1 =    1.8600
           m2 =    2.2100
          sd1 =    1.2900
          sd2 =    1.8000

Estimated sample sizes:

            N =       498
  N per group =       249
```

We see that the required sample size is 498, with 249 in each group. The principal reason why this required sample size is so high is that the assumed effect size (that is, the assumed difference between the two population means) is relatively small. Detecting a smaller effect size requires a larger sample.

The graph option can again be used. This time we obtain a plot of the required sample size against power, using the command:

```
power twomeans 1.86 2.21 , sd1(1.29) sd2(1.80) oneside power(0.1(0.1)0.9) graph
```

The result is shown in Figure 6'.

1.4.3 Power analysis for tests of equality of variance

Another test that was performed in Section 1.2.2 was the equality of variances test, `sdtest`. The results were as follows:

```
. sdtest v_mug, by(v_type)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTP |      49    1.862245     .184379    1.290653    1.491526    2.232964
     WTA |      51    2.213725    .2515679    1.796554    1.708436    2.719015
---------+--------------------------------------------------------------------
combined |     100      2.0415    .1571287    1.571287    1.729723    2.353277
------------------------------------------------------------------------------
    ratio = sd(WTP) / sd(WTA)                                     f =   0.5161
Ho: ratio = 1                                 degrees of freedom =   48, 50

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0114         2*Pr(F < f) = 0.0229           Pr(F > f) = 0.9886
```

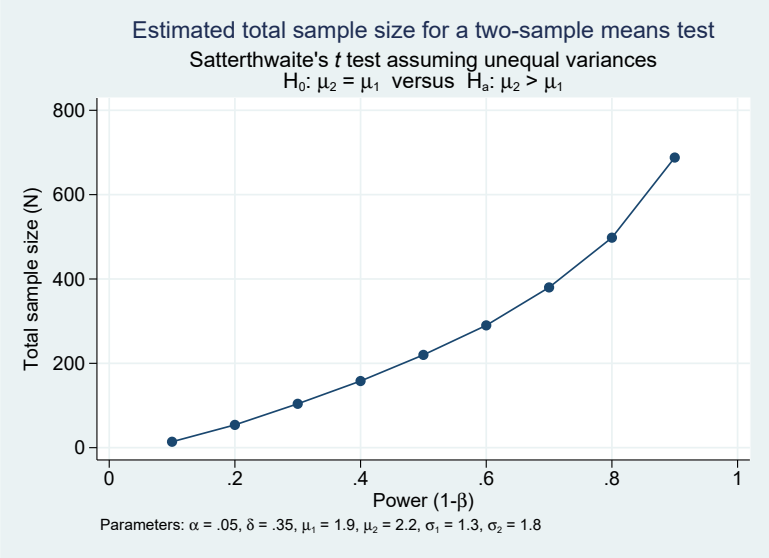


Figure 6: Required sample size against power

We found a significant difference (p=0.0229 2-tailed) between the variances of WTP and WTA.

To apply power analysis to this test, we first need the two variances. These are $V(WTP) = 1.29^2 = 1.66$ and $V(WTA) = 1.80^2 = 3.24$. We then apply the `power` command as follows:

```
. power twovariances 1.66 3.24, n1(49) n2(51)

Estimated power for a two-sample variances test
F test
Ho: v2 = v1  versus  Ha: v2 != v1

Study parameters:

        alpha =    0.0500
            N =       100
           N1 =        49
           N2 =        51
        N2/N1 =    1.0408
        delta =    1.9518
           v1 =    1.6600
           v2 =    3.2400

Estimated power:

        power =    0.6402
```

As usual, it seems that we require a larger sample in order to achive the desired power of 0.8. To find this sample size, we use:

```
. power twovariances 1.66 3.24, p(0.8)

Performing iteration ...

Estimated sample sizes for a two-sample variances test
F test
Ho: v2 = v1  versus  Ha: v2 != v1

Study parameters:

        alpha =    0.0500
```

```
        power =    0.8000
        delta =    1.9518
           v1 =    1.6600
           v2 =    3.2400

Estimated sample sizes:

            N =       146
  N per group =        73
```

It seems that a sample of 73 per group is required for this test.

1.4.4 Power analysis for the paired test on valuation data

In Section 1.2.6 we reported the following results of a paired test comparing WTA and WTP, for a lottery, for a sample of 100 subjects:

```
. ttest WTA=WTP

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     WTA |     100      2.5745    .0961546    .9615459    2.383708    2.765292
     WTP |     100      2.2415    .1115095    1.115095    2.020241    2.462759
---------+--------------------------------------------------------------------
    diff |     100        .333    .1398705    1.398705    .0554666    .6105334
------------------------------------------------------------------------------
     mean(diff) = mean(WTA - WTP)                                     t =   2.3808
 Ho: mean(diff) = 0                                     degrees of freedom =       99

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9904         Pr(|T| > |t|) = 0.0192          Pr(T > t) = 0.0096
```

We saw that there is strong evidence (p-value=0.0096) of WTA being greater than WTP.

Let us now use the `power` command to compute the power of the test just performed.

```
. power pairedmeans 2.57 2.24 , sd1(0.96) sd2(1.11) corr(0) n(100) oneside

Estimated power for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d < d0

Study parameters:

        alpha =    0.0500          ma1 =    2.5700
            N =       100          ma2 =    2.2400
        delta =   -0.2249          sd1 =    0.9600
           d0 =    0.0000          sd2 =    1.1100
           da =   -0.3300         corr =    0.0000
         sd_d =    1.4675

Estimated power:

        power =    0.7219
```

The power is 0.72, indicating that the sample size needs to be somewhat greater than 100 to obtain a power of 0.80. The required sample size is found as follows:

```
. power pairedmeans 2.57 2.24 , sd1(0.96) sd2(1.11) corr(0) p(0.80) oneside

Performing iteration ...

Estimated sample size for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d < d0
```

```
Study parameters:

        alpha =    0.0500          ma1 =    2.5700
        power =    0.8000          ma2 =    2.2400
        delta =   -0.2249          sd1 =    0.9600
           d0 =    0.0000          sd2 =    1.1100
           da =   -0.3300         corr =    0.0000
         sd_d =    1.4675

Estimated sample size:

            N =       124
```

The required sample is 124. However, note that we have used the `corr(0)` option. It plausible to expect the correlation between WTP and WTA to be positive. Returning to the data, we find this correlation as follows:

```
. corr WTA WTP
(obs=100)

             |      WTA      WTP
-------------+------------------
         WTA |   1.0000
         WTP |   0.0987   1.0000
```

The correlation is found to be +0.10, and repeating the sample-size calculation assuming this correlation gives:

```
. power pairedmeans 2.57 2.24 , sd1(0.96) sd2(1.11) corr(0.10) p(0.80) oneside

Performing iteration ...

Estimated sample size for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d < d0

Study parameters:

        alpha =    0.0500          ma1 =    2.5700
        power =    0.8000          ma2 =    2.2400
        delta =   -0.2369          sd1 =    0.9600
           d0 =    0.0000          sd2 =    1.1100
           da =   -0.3300         corr =    0.1000
         sd_d =    1.3930

Estimated sample size:

            N =       112
```

The required sample size is now 112.

2 Power Analysis using Monte Carlo

2.1 The Monte Carlo Method

If you want to be convinced that a particular test does what it is supposed to do, or if you want to consider which of a number of tests performs best in a particular situation, the Monte Carlo Method is very useful.

Note that the STATA `power` command, demonstrated in previous sections, is very useful for carrying out power calculations for simple parametric tests, particularly tests based on the t-distribution. If we require to find the power of a parametric treatment test in a complex model, or the power of a non-parametric test, the Monte Carlo method is required.

2.1.1 Finding (actual) size and power of tests using the Monte Carlo method

In do-file 2, there is an example of the use of the `simulate` command - the Monte Carlo command in STATA. In the example, we compare the performance (size and power) of three tests under different assumptions. The three tests are:

1. The independent samples t-test. `Compares two means`
2. The Mann-Whitney test `Compares two medians`
3. The Kolmogorov-Smirnov test `Compares two distributions`

The simulation is based on the following simple data generating process:

$$\begin{aligned} x_i &= 10 + \delta d_i + \epsilon_i \quad i = 1, \ldots, n \\ \texttt{Control} \quad & d_i = 0 \text{ if } i \leq n/2 \\ \texttt{Treatment} \quad & d_i = 1 \text{ if } i > n/2 \\ & V(\epsilon_i) = 1 \end{aligned} \tag{16}$$

In (16), d_i is a dummy variable representing treatment: 1 for treatment; 0 for control. The first half of the sample is control; the second half is treatment. The parameter δ is the treatment effect.

The first thing to do is to find the *actual size* of each test, that is, the proportion of replications for which the hypothesis $\delta = 0$ is rejected *when the true value of δ is zero*. An important requirement of a test is that the actual size is close to the nominal size (typically 0.05). We then find the power when the treatment effect is $\delta = 0.5$.

Size and power of the three tests are shown in the table below. These numbers are from a Monte Carlo with 1000 replications, with the assumption of normality of the error term ϵ_i.

	SIZE	POWER (δ=0.5)
t-test	0.052^u	0.702
MW	0.053^u	0.683
KS	0.040^u	0.513

A u-superscript in the first column indicates that the proportion of rejections is *not* significantly different from 0.05, implying that the test is correctly sized or *unbiased* (Feiveson et al. 2002). We see that all three tests are correctly sized, so we may compare them on the criterion of power.

We see that the independent samples t-test is the most powerful of the three tests, and the Kolmogorov-Smirnov test is the least powerful.

The power of 0.702 for the t-test is of course something that can be computed using the `power` command. Let us check that this gives the same answer:

```
. power twomeans 10 10.5, sd(1) n(100)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =       100
  N per group =        50
        delta =    0.5000
           m1 =   10.0000
           m2 =   10.5000
           sd =    1.0000

Estimated power:

        power =    0.6969
```

It is straightforward to check that the proportion 0.702 obtained from the Monte Carlo is not significantly different from the 0.697 obtained from the `power` command.

The superiority of the t-test may be simply a result of the assumptions underlying the t-test being met. Let us repeat the process with different assumptions about the distribution of the error term.

If the error term ϵ_i in (16) is assumed to be $U(-2, 2)$, we obtain the size and power shown in the following table.

	SIZE	POWER (δ=0.5)
t-test	0.056^u	0.566
MW	0.056^u	0.526
KS	0.039^u	0.306

We see that, again, all three tests are correctly sized, and again the t-test is the most powerful and the KS test least powerful.

The third distribution we try is a skewed distribution. The error term is a $\chi^2(3)$ distribution standardised to have mean zero and variance 1. The results are shown in the next table.

	SIZE	POWER (δ=0.5)
t-test	0.061^u	0.705
MW	0.067	0.867
KS	0.052^u	0.862

This time, we see that the MW test is excessively sized, so we need to discard it. The other two tests are correctly sized, and of these two, the KS test is much more powerful.

The message seems to be that we should worry about using conventional tests whenever the distribution of the data is highly asymmetric.

```
Q: What if the distribution is skewed toward the opposite
direction?
A: We can make it minus the error term in the do-file.
However, we run the simulation and find different results
(Mann-Whitney doing better than K-S  test.)

Q: Can we try the E-S test?
A: Yes, and it is in "do-file-2a"!
```

2.2 Treatment testing with multi-level data

Experimental data rarely consists of independent observations. There is often dependence at different levels. For example, if each subject makes a sequence of decisions, there is dependence at the level of individual subjects. In interactive experiments, there is also likely to be dependence at the level of the group of subjects, or at the level of the session in which the groups of subjects perform their tasks.

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

In do-file 2b, there is a Monte-Carlo program that simulates data from an experiment with both subject level and group-level clustering. We are once again interested in tests of a treatment effect. Seven different testing procedures are used. Each is a t-test from a particular regression model. The seven models are:

1. OLS no clustering
2. OLS with clustering at the subject level
3. OLS with clustering at the group level
4. Random effects, no clustering
5. Random effects, with clustering at the subject level
6. Random effects, with clustering at the group level
7. Multi-level model (subject random effect and group random effect)

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

The questions we set out to answer with the Monte Carlo are:

1. Which of these testing methods are correctly sized?
2. Of those which are correctly sized, which has highest power?

2.2.1 The multi-level model

First we will explain the structure of the multi-level model. Note that the other models (ols and random effects) can be seen as special cases of this general model.

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. 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 groups, a typical observation is represented by y_{ijt}, and a "three-level" model is appropriate, with the group indicator j representing the third (or "highest") level of clustering.

Subject random effect

The three-level model just described might be specified as follows.

Group random effect

d is the treatment, say auction format

$$
\begin{aligned}
y_{ijt} &= \alpha + \delta d_i + \beta x_{ijt} + u_i + v_j + \epsilon_{ijt} \\
i = 1 \ldots, n \quad j = 1 \ldots, J \quad t = 1 \ldots, T & \\
Var(u_i) &= \sigma_u^2 \\
Var(v_j) &= \sigma_v^2 \\
Var(\epsilon_{it}) &= \sigma_\epsilon^2
\end{aligned} \tag{17}
$$

STATA: xtmixed

In (17), y_{ijt} might be the bid of subject i in group j in round t in an auction or contest. The variable x_{ijt} might be the private signal received by subject i in round t. d_i is a treatment dummy. The treatment might be a "low uncertainty" treatment, in which there is lower uncertainty over the object of the bidding, and we might expect bids to be higher in this treatment, that is, we would expect the treatment effect δ to be positive. u_i is the subject-specific random effect, v_j is the session-specific random effect, and ϵ_{ijt} is the observation-specific error term.

Notice that if you remove v_j from (17), you have the random effects model. If you remove both v_j and u_i, you have a linear regression model.

In (17), the treatment dummy (d_i) has only an i subscript. This implies that the treatment is being applied between-subject: some subjects are exposed to the treatment throughout the experiment; others are not. Of course, it would be possible to apply the treatment within-subject, that is, for all subjects to experience the treatment for (say) half of the tasks. In this case, the treatment variable appearing in (17) would be d_{it}, that is it would have both i and t subscripts.

The STATA command for multi-level modelling is `xtmixed`. An example of the use of the command is:

Cluster at two levels

```
. xtmixed y d x   || j: || i:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -2959.3982
Iteration 1:   log likelihood = -2959.3978
Iteration 2:   log likelihood = -2959.3978

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =      2,000

------------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
```

```
---------------+-----------------------------------------
             j |         10         200       200.0         200
             i |         40          50        50.0          50
---------------------------------------------------------------

                                                Wald chi2(2)      =     155.37
Log likelihood = -2959.3978                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           d |   .1482739   .0454989     3.26   0.001     .0590978      .23745
           x |   .0955655   .0079035    12.09   0.000     .0800749     .111056
       _cons |  -.1241784    .247917    -0.50   0.616    -.6100867     .3617299
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
j: Identity                  |
                sd(_cons) |   .4820359    .292011      .1470391    1.580251
-----------------------------+------------------------------------------------
i: Identity                  |
                sd(_cons) |   1.193918    .156372      .9236118    1.543333
-----------------------------+------------------------------------------------
             sd(Residual) |   1.017198   .0162466      .9858481    1.049544
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 1737.24                   Prob > chi2 = 0.0000

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

2.2.2 Results from the Monte Carlo study

We assume $n = 40$ subjects and $T = 50$ rounds. We also assume that subjects are divided into groups of 4.

```
egen i=seq(), f(1) b(50) * from 1 by 50 means
(1,...,1,2...,2,...)
egen j=seq(), f(1) t(50) * from 1 to 50 means
(1,23,4,...50,1,2,...,50,...)
```

We use 100 replications in the Monte Carlo.

The following table shows the results for the between-subject treatment tests (in which half of the subjects are exposed to the treatment). The treatment effect under the alternative hypothesis is $\delta = 0.5$.

d=0

	SIZE	POWER (δ=0.5)
ols no clustering	0.46	0.68
OLS with clustering at the subject level	0.15	0.41
OLS with clustering at the group level	0.07[u]	0.25
Random effects no clustering	0.13	0.41
Random effects with clustering at the subject level	0.15	0.41
Random effects with clustering at the group level	0.07[u]	0.25
Multi-level model	0.08[u]	0.27

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

Interestingly, the clustering models that result in unbiased tests are the ones that deal with clustering at the group level. Dealing with clustering at the lower level of

Need to cluster at the highest level!

the individual appears to be inadequate.

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

d=t>T/2 instead of d=i>N/2

Next, let us turn to the within-subject tests (in which all subjects experience the treatment in half of the rounds. Since within-subject tests can detect smaller treatment effects, we shall assume a much smaller treatment effect of $\delta = 0.05$ under the alternative hypothesis. The results are as follows.

d=0

	SIZE	POWER (δ=0.05)
ols no clustering	0.02^u	0.07
OLS with clustering at the subject level	0.09^u	0.31
OLS with clustering at the group level	0.09^u	0.33
Random effects no clustering	0.05^u	0.31
Random effects with clustering at the subject level	0.09^u	0.31
Random effects with clustering at the group level	0.08^u	0.33
Multi-level model	0.05^u	0.31

We see very different results from the between-subject tests. All seven tests appear to be unbiased (although if we used a larger number of replications in the Monte-Carlo, we are likely to find that some are incorrectly sized). The testing procedure that ignores clustering has very low power. The other six have modest power, and there is very little difference between them.

2.2.3 Summary of results

The key results of this section are: in the between-subject context, only three of the seven tests are correctly sized: the two that use group-level clustering, and the multi-level model; of these three tests, the most powerful is the one performed in the framework of the multi-level modelling. Failure to deal with clustering has very serious consequences in terms of massively excessive test size.

Recommendations that follow from these results are: in the between-subject context, the multi-level model is the best model in which to conduct treatment tests; if clustering is to be used, it is preferable to cluster at the highest possible level (e.g. group rather than subject).

In the within-subject context, the results are very different. Firstly, within-subject tests are able to detect much smaller treatments than within-subject tests. All of the approaches perform well on both size and power, except ols without clustering.

What if we make the group effect much smaller? 0.1 instead of 1 makes the size of other regressions also good since group effects is now small.
(gen y=0.5+delta*d+0.1*x+u+0.1*v+e)

However, power has little change. More research needed here.

2.3 Varying n and T

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

Clearly an increase in n and an increase in T can both be expected to increase the power of the treatment test. But which of the two is more beneficial? Should one be increased at the expense of the other? We shall attempt to answer these questions.

2.3.1 The effect of increasing n and T on power in the multi-level model

do-file_2c contains the code for the Monte Carlo. `J = 10 groups`
`Typo: ‘ in wrong place for STATA command gen d=i/2`
The following table shows the results for the between-subject tests, with a treatment effect of $\delta = 0.5$. The numbers shown in the table are powers for different combinations of n and T.

	$T = 50$	$T = 100$	$T = 150$
$n = 40$	0.24	0.26	0.28
$n = 80$	0.25	0.34	0.35
$n = 120$	0.39	0.38	0.35

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

The following table shows the results for the within-subject tests, with a smaller treatment effect of $\delta = 0.05$.

	$T = 50$	$T = 100$	$T = 150$
$n = 40$	0.20	0.47	0.75
$n = 80$	0.44	0.71	0.91
$n = 120$	0.67	0.81	0.97

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

3 Estimation of risk aversion parameters using risky choice data

3.1 Modelling choices between lotteries (the "house money effect")

In this sub-section, we consider a very popular application of binary data models: risky choice experiments. We will use the models to test a particular hypothesis relating to behaviour in this context. The hypothesis of interest is the "house money effect", that is, the phenomenon of choices becoming more risk-seeking when the initial endowment is higher (see Thaler & Johnson 1990, Keasey & Moon 1996).

Consider the choice problem presented in Figure 7, where the two circles represent lotteries, and the areas within them represent probabilities of the stated outcomes (~~the same lottery choice example was used to demonstrate the use of non-parametric tests in Part 1~~. 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.

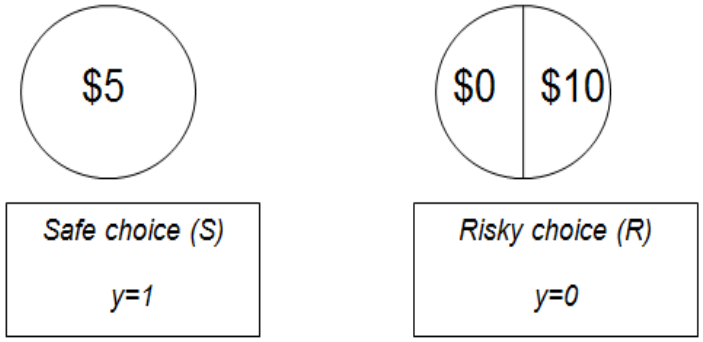


Figure 7: A lottery choice problem

Clearly, by choosing between the lotteries in Figure 7, a subject is conveying some information about his or her attitude to risk. What is of interest here is whether previously endowing a subject with an amount of money has an effect on this choice. Let us define the "house money effect" as the phenomenon of agents becoming less risk averse (i.e. more likely to choose the risky lottery) when their initial endowment (i.e. "house money") increases.

Suppose we have a sample of 1,050 subjects. We endow each subject (i) with a different wealth level (w_i); we then immediately ask them to choose between the two lotteries shown in Figure 7. We then define the binary variable y to take the value 1 if the safe lottery is chosen, and 0 if risky is chosen. The results of this (imaginary) experiment are contained in the file **house_money_sim**. Here is some summary information about the data:

```
. table w, contents(n y mean y)
--------------------------------
       w |       N(y)     mean(y)
```

0	50	.92
.5	50	.88
1	50	.88
1.5	50	.84
2	50	.84
2.5	50	.9
3	50	.84
3.5	50	.72
4	50	.78
4.5	50	.7
5	50	.7
5.5	50	.74
6	50	.72
6.5	50	.72
7	50	.5
7.5	50	.64
8	50	.5
8.5	50	.48
9	50	.56
9.5	50	.5
10	50	.5

The final column of the table shows the mean of the binary variable for different wealth levels. Since the mean of a binary variable is the proportion of ones in the sample, the numbers in this column represent the proportion choosing the safe lottery at each wealth level. The tendency for this proportion to fall as wealth rises is consistent with the house money effect.

Next we set out to confirm this using a parametric model. A natural model to start with is the probit model, defined as follows:

$$P\left(y_i = 1|w_i\right) = \Phi\left(\beta_0 + \beta_1 w_i\right) \tag{18}$$

where $\Phi(.)$ is the standard normal c.d.f.[7] The likelihood function for the probit model is:

$$L = \prod_{i=1}^{n} \left[\Phi\left(\beta_0 + \beta_1 w_i\right)\right]^{y_i} \left[1 - \Phi\left(\beta_0 + \beta_1 w_i\right)\right]^{1-y_i} \tag{19}$$

and the log-likelihood is:

Why take log? Because the likelihood is a bunch of probability multiplied together, so as n increases, L essentially goes to zero and is difficult to maximize.

$$LogL = \sum_{i=1}^{n} \left[y_i ln\left(\Phi\left(\beta_0 + \beta_1 w_i\right)\right) + \left(1 - y_i\right) ln\left(1 - \Phi\left(\beta_0 + \beta_1 w_i\right)\right)\right] \tag{20}$$

An important property of the cdf (18) defining the probit model is symmetry. By this, we mean that $\Phi(-z) = 1 - \Phi(z)$. This property also applies to the distribution underlying the logit model (see Exercise 1). This feature of the underlying distribution is useful because it allows the log-likelihood function to be written more compactly as follows. If we recode the binary variable as:

$yy_i = \quad 1$ if S is chosen
$yy_i = -1$ if R is chosen

[7] If a random variable Z has a standard normal distrubution, its density function is $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$, and its cumulative distribution function (c.d.f.) is $\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \phi(z)\mathrm{d}z$.

then the log-likelihood (20) can be written as:

$$LogL = \sum_{i=1}^{n} ln\left(\Phi\left(yy_i \times (\beta_0 + \beta_1 w_i)\right)\right) \tag{21}$$

We maximise LogL defined in (21) to give MLE's of the two parameters β_0 and β_1. This task is performed using the `probit` command in STATA, as follows:

```
. probit y w

Iteration 0:   log likelihood =  -634.4833
Iteration 1:   log likelihood = -584.91375
Iteration 2:   log likelihood =  -584.5851
Iteration 3:   log likelihood = -584.58503
Iteration 4:   log likelihood = -584.58503

Probit regression                                 Number of obs   =       1050
                                                  LR chi2(1)      =      99.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -584.58503                       Pseudo R2       =     0.0786

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |  -.1409882   .0145377    -9.70   0.000    -.1694816   -.1124948
       _cons |   1.301654   .0911155    14.29   0.000     1.123071    1.480237
------------------------------------------------------------------------------
```

The first thing we might do when we have obtained the results is to test for the presence of the house money effect. This test is done for us. The asymptotic t-statistic associated with wealth is $z = -9.70$, and the associated p-value is 0.000. This tells us that there is strong evidence that wealth has a negative effect on the probability of choosing the safe lottery. In other words, there is strong evidence of the house money effect in this data.

There is a STATA command `test` that can be used immediately after estimation of a model. Using this for the test just performed, we obtain:

```
. test w=0

 ( 1)  [y]w = 0

           chi2(  1) =   94.05
         Prob > chi2 =    0.0000
```

This is a Wald test of the house money effect. The Wald test statistic is the square of the asymptotic t-test statistic $[94.05 = (-9.70)^2]$, and has a $\chi^2(1)$ distribution under the null hypothesis of no effect. The Wald test is equivalent to the asymptotic t-test and the two tests will always have the same p-value.

The next thing we might wish to do is to predict the probability of making the safe choice at each wealth level. The best way of presenting this is using a graph of predicted probability against w. The formula we need to graph is $\Phi(1.302 - 0.141w)$. The graph can be obtained using the following two STATA commands:

```
margins, at(w=(0(1)15))
marginsplot, ylabel(0(0.1)1) yline(0.5)
```

The result is shown in Figure 8. We see that when the initial endowment is 0, there is a high probability that the safe alternative will be chosen, that is, subjects appear to be highly risk averse. We also see that as the initial endowment rises, the

probability of choosing the safe alternative falls fairly steeply. Since a probability of 0.5 is associated with risk-neutrality, and, remembering that $\Phi^{-1}(0.5) = 0$, it appears that, in order to induce risk-neutrality in subjects, it is necessary to endow them with an amount $1.3016/0.1410 = \$9.23$. When the initial endowment is above this amount, risk-seeking behaviour is predicted, since the predicted probability of the safe choice is then lower than 0.5.

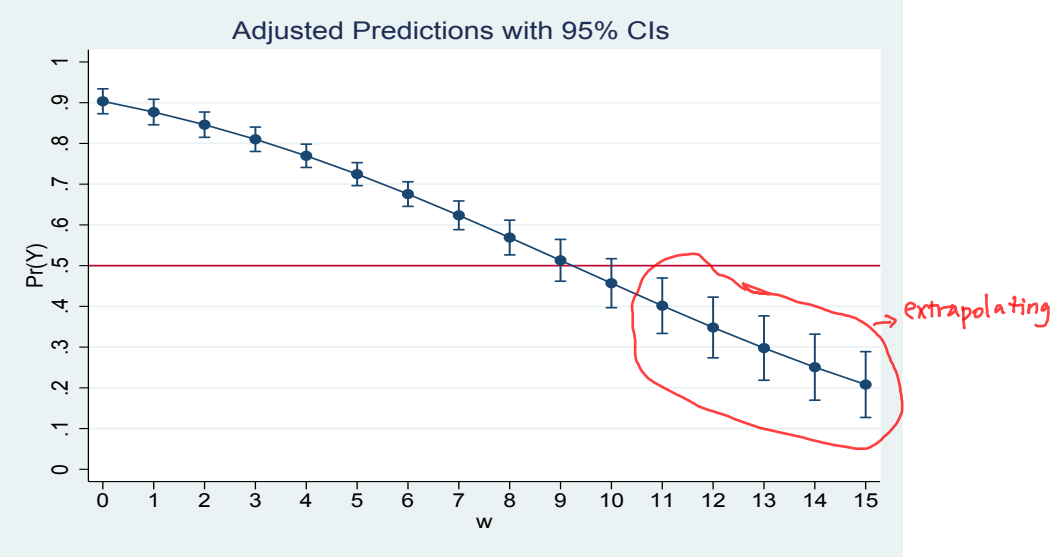


Figure 8: Predicted probabilities of safe choice against wealth level from the probit model.

3.1.1 Marginal effects

Something else that is sometimes useful after estimating a probit model is to obtain conditional marginal effects. This is the predicted change in the probability resulting from a small change in the explanatory variable starting from a particular value. For example, if we wish to know how much the probability of S changes when w rises from 0, we use:

```
. margins, dydx(w) at(w=0)

Conditional marginal effects                      Number of obs   =       1050
Model VCE    : OIM

Expression   : Pr(y), predict()
dy/dx w.r.t. : w
at           : w               =           0

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |   -.024109   .0013299   -18.13   0.000    -.0267155   -.0215026
------------------------------------------------------------------------------
```

We see that the conditional marginal effect is -0.024, implying that, roughly speaking, if w rises from 0 to 1, the probability of S will fall by 2.4 percentage points. If we condition on a higher value of w, we obtain a different result:

```
. margins, dydx(w) at(w=10)

Conditional marginal effects                      Number of obs   =       1050
```

```
Model VCE    : OIM

Expression   : Pr(y), predict()
dy/dx w.r.t. : w
at           : w               =          10

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |  -.0559177   .0053804   -10.39   0.000    -.0664631   -.0453724
------------------------------------------------------------------------------
```

This higher (in magnitude) marginal effect (-0.056) simply reflects the fact that the curve shown in Figure 8 is steeper at w=10 than at w=0. Finally, if we use the `margins` command without the `at( )` option, we obtain the *average marginal effect*.

```
. margins, dydx(w)

Average marginal effects                        Number of obs   =       1050
Model VCE    : OIM

Expression   : Pr(y), predict()
dy/dx w.r.t. : w

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |  -.0444259   .0039929   -11.13   0.000    -.0522518      -.0366
------------------------------------------------------------------------------
```

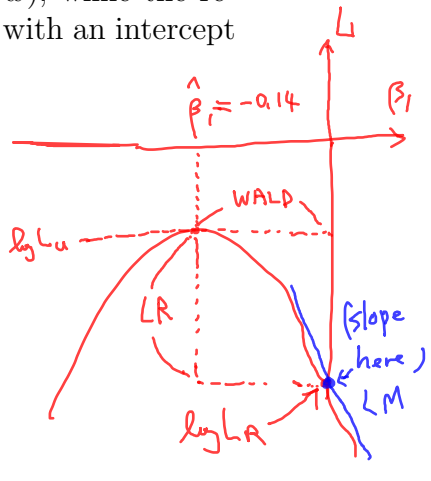
The average marginal effect is seen to be -0.044. This is simply the average of the marginal effects over all of the observations in the sample.

3.1.2 Wald tests and LR tests

The test used in Section 3.1 for the significance of the variable w was a Wald test. It was demonstrated that this test can be conducted using the `test` command, and the Wald test statistic is the square of the asymptotic t-test statistic.
There is yet another way of testing the same hypothesis: the likelihood ratio (LR) test. This test is based on a comparison of the maximised log-likelihood in two different models. The test statistic is computed using:

$$LR = 2(LogL_U - LogL_R) \tag{22}$$

where $LogL_U$ is the maximised log-likelihood from the unrestricted model, and $LogL_R$ is the same for the restricted model. In the present case, the unrestricted model is the model that has been estimated (probit model with w), while the restricted model is a probit model with w removed, that is, a model with an intercept only. Estimation of this restricted model gives:




```
. probit y

Iteration 0:   log likelihood =  -634.4833
Iteration 1:   log likelihood =  -634.4833

Probit regression                                 Number of obs   =       1050
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood =  -634.4833                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5464424   .0408516    13.38   0.000     .4663746    .6265101
------------------------------------------------------------------------------
```

We see that the restricted log-likelihood is −634.48. This is used to compute the LR test statistic using (22):

$$LR = 2(LogL_U - LogL_R) = 2(-584.59 - (-634.48)) = 99.8 \tag{23}$$

Under the null hypothesis of no house money effect, the statistic given by (23) comes from a $\chi^2(1)$ distribution. Therefore we reject the null because $99.8 > \chi^2_{1,0.05} = 3.84$. In fact, there is a way of computing the LR test statistic directly in STATA. The estimates from the two models are stored, and then the **lrtest** command is applied. The required sequence of commands, and the results, are as follows:

```
probit y w
est store with_w

probit y
est store without_w

lrtest with_w without_w

Likelihood-ratio test                                 LR chi2(1)  =     99.80
(Assumption: without_w nested in with_w)              Prob > chi2 =    0.0000
```

Reassuringly the result is exactly the same as (23). An advantage of using STATA to perform the test is that a p-value is provided in addition to the test statistic. In this case the p-value (0.0000) conveys overwhelming evidence of the house money effect.

Finally, note that the LR test statistic (99.80) is fairly close to the Wald test statistic (94.05) for the same hypothesis. This similarity is not surprising since the two tests are asymptotically equaivalent.

3.2 Analysis of ultimatum game data

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.

~~In Section ??, we analysed the proposers' offers in this data set, and we tested for a gender effect~~. In this section, we will turn to the responder's decision. This is a binary decision, so binary data models are required to identify its determinants.

We first consider simply how many of the subjects rejected offers. For this we obtain a tabulation of the binary variable, from which we see that 51 of the 200 subjects (approximately one-quarter of them) rejected offers.

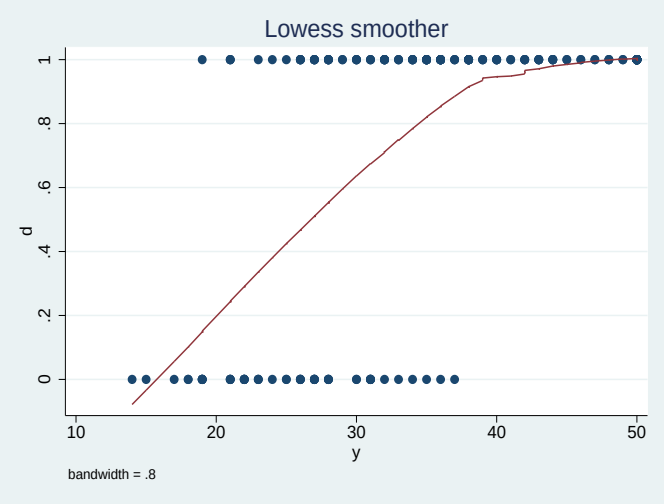


Figure 9: Responder's decision (d) against proposer's offer (y), with smoother

```
. tab d

          d |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         51       25.50       25.50
          1 |        149       74.50      100.00
------------+-----------------------------------
      Total |        200      100.00
```

To show overlapping points, use:
lowess d y jitter(5) msize(3)

The main determinant of the responder's decision is the proposer's offer (y). Sometimes it is useful to plot binary data. The command `lowess d y` produces the graph shown in Figure 9. Lowess (locally weighted scatter-plot smoother) is a form of non-parametric regression that is used elsewhere in the course. Roughly speaking, it shows the mean value of d conditional on different values of y. Since the mean of d is closely related to the probability of the offer being accepted, the graph is telling us that the probability of acceptance rises sharply as the offer rises, approaching 1 as the offer approaches 50.

In complete contrast to "Lowess", the probit model introduced in Section 3.1 is an example of a fully parametric estimation procedure. The probit model is defined as follows:

$$P(d = 1|y) = \Phi(\beta_0 + \beta_1 y) \tag{24}$$

where $\Phi(.)$ is the standard normal cdf. The results are as follows:

```
. probit d y

Iteration 0:   log likelihood = -113.55237
Iteration 1:   log likelihood = -70.230335
Iteration 2:   log likelihood = -66.806698
Iteration 3:   log likelihood = -66.738058
Iteration 4:   log likelihood = -66.738049
Iteration 5:   log likelihood = -66.738049

Probit regression                                 Number of obs   =        200
                                                  LR chi2(1)      =      93.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -66.738049                       Pseudo R2       =     0.4123

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |   .1439157   .0212804     6.76   0.000     .1022069    .1856244
```

```
       _cons |  -3.855266    .631443    -6.11   0.000    -5.092872   -2.617661
------------------------------------------------------------------------------
```

From the results, we can deduce a formula for the predicted probability of an offer (y) being accepted:

$$\hat{P}(d=1|y) = \Phi(-3.855 + 0.144y) \tag{25}$$

In this situation, it is useful to consider the equation *underlying* the probit model:

$$\begin{aligned} d^* &= \beta_0 + \beta_1 y + \epsilon \\ \epsilon &\sim N(0,1) \end{aligned} \tag{26}$$

like "utility"

In (26), d^* is the propensity of the responder to accept the offer. If this propensity is greater than zero, the offer is accepted:

$$d = 1 \Leftrightarrow d^* > 0 \Leftrightarrow \beta_0 + \beta_1 y + \epsilon > 0 \Leftrightarrow \epsilon > -\beta_0 - \beta_1 y \tag{27}$$

Hence the probability of the offer being accepted is:

$$P(d=1) = P(\epsilon > -\beta_0 - \beta_1 y) = \Phi(\beta_0 + \beta_1 y) \tag{28}$$

which is the probability formula (24) on which the probit model is based. The reason why (26) is useful is because it enables us to compute the "minimum acceptable offer (MAO)" for a typical subject. Disregarding the error term, we have:

$$d^* = \beta_0 + \beta_1 y \tag{29}$$

A typical subject is indifferent between accepting and rejecting an offer when (29) is zero:

$$\beta_0 + \beta_1 y = 0 \Rightarrow y = -\frac{\beta_0}{\beta_1} \tag{30}$$

We compute this from the estimates as follows:

$$y^{MAO} = -\frac{-3.855}{0.144} = 26.79 \tag{31}$$

The MAO (31) can also be computed in STATA with the nlcom command:

```
. nlcom MAO: -_b[_cons]/_b[y]

        MAO:  -_b[_cons]/_b[y]

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         MAO |   26.78837   .9268278    28.90   0.000     24.97182    28.60492
------------------------------------------------------------------------------
```

The `nlcom` procedure uses a technique known as the "delta method" which is considered in more detail in Section 3.5. The major benefit from applying this technique is that it returns a standard error and confidence interval for MAO, in addition to the point estimate. The point estimate of 26.79 is telling us that a "typical" responder (typical in the sense of having an error term ϵ equal to the mean value of 0) would say "no" to the offer of 26, but "yes" to 27.

3.2.1 The strategy method

The strategy method has been used in the context of the ultimatum game by Solnick (2001), among others. The proposer makes an offer as before. Let this offer be y. Meanwhile, in a different room, the responder is asked to state their minimum acceptable offer (y^{MAO}). Then y is compared to y^{MAO}. If $y \geq y^{MAO}$, the offer is taken as being accepted, and both players receive their pay-offs. If $y < y^{MAO}$, the offer is taken as being rejected, and both players receive zero.

Under this approach, the responder is not only being asked for a decision, but for their *strategy*. Note that it is in their interest to state their MAO truthfully; for this reason, the strategy method is said to be *incentive compatible*.

The standard version of the ultimatum game (~~i.e. as described in Section **??**~~) is known as the "direct decision apprach". The strategy method has a considerable advantage over the direct decision approach. The data is much more informative. Clearly, it is more useful to know the responder's minimum acceptable offer than it is simply to know whether they have accepted a particular offer. This is particularly so in the cases where proposers offer 50% of their endowment. When a proposer offers 50%, the offer is almost certain to be accepted, and very little is learned, despite a significant cost to the experimenter. The strategy method enables useful information to be learned from all responders.

Let us imagine that the strategy method has been applied to the 200 subjects instead of the "direct decision approach", and that the data set consists of:

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;
MAO:	responder's minimum acceptable offer;
d:	outcome: 1 if $y \geq y^{MAO}$; 0 if $y < y^{MAO}$.

For earlier versions of STATA: ci MAO

A (simulated) data set containing these variables is contained in the file **ug_sm_sim**. With this data, we carry out the following simple analysis:

```
. ci means MAO

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         MAO |        200      31.375    .6666664        30.06036    32.68964

. tab d

          d |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         87       43.50       43.50
          1 |        113       56.50      100.00
------------+-----------------------------------
      Total |        200      100.00
```

The straightforward command `ci MAO` has been used to obtain a 95% confidence interval for the population mean of y^{MAO}. This confidence interval is clearly narrower than the one obtained in Section 3.2.1 from the direct-decision data (24.97182

→ 28.60492). This simply confirms that we are able to estimate parameters more precisely when the strategy method has been used.

However, note also that the MAO appears to be around five units higher when the strategy method is being used. This has the consequence that the number of "rejections" is higher (87, compared with 51 under the direct-decision approach). This is a common finding. Eckel & Grossman (2001) explain the higher rejection rate under the strategy method in terms of subjects' failure to understand the simultaneous nature of the decision and their attempt to signal a "tough" bargaining position.

There may also be a perception on the part of the responder that their statement of MAO is *hypothetical* (even though it is not, since it determines their pay-off). Asking a subject about how they *would* act in different situations is sometimes referred to as a "cold" treatment, to be contrasted with the "hot" treatment that arises when an offer is actually placed in front of the responder, and all they need to do is "accept".

The message here is perhaps that superior data should be obtained using the strategy method, but that an adjustment should be applied to the MAO data in order for it to be applicable to the "direct decision" situation. On the evidence above, the stated values of MAO would need to be reduced by around five units.

3.3 The `ml` Routine in STATA

There is another way to estimate the probit model. This is to specify the log-likelihood function ourselves, and ask STATA to maximise it. We shall return to the **house_money_sim** data set used in Section 3.1.

The following code defines a program called `myprobit` which computes the log-likelihood. It then reads the data and calls on the `ml` program to perform the maximisation of the log-likelihood. The formula being programmed is the one given in (21) above:

capture program drop myprobit
"Capture" ignores error if the following command is nor applicable.

$$LogL = \sum_{i=1}^{n} ln\left(\Phi\left(yy_i \times (\beta_0 + \beta_1 w_i)\right)\right) \tag{32}$$

```
* LOG-LIKELIHOOD EVALUATION PROGRAM "myprobit" STARTS HERE:

program define myprobit

* SPECIFY NAME OF QUANTITY WHOSE SUM WE WISH TO MAXIMISE (logl)
* AND ALSO PARAMETER NAMES (EMBODIED IN xb)
* PROVIDE LIST OF TEMPORARY VARIABLES (p ONLY)

args logl xb
tempvar p

* GENERATE PROBABILITY OF CHOICE MADE BY EACH SUBJECT (p):

quietly gen double `p'=normal(yy*`xb')

* TAKE NATURAL LOG OF p AND STORE THIS AS logl

quietly replace `logl'=ln(`p')

* END "myprobit" PROGRAM:
```

```
end

* READ DATA

use "house money_sim", clear

* GENERATE (INTEGER) yy FROM y:

gen int yy=2*y-1

* SPECIFY LIKELIHOOD EVALUATOR (lf), EVALUATION PROGRAM (myprobit),
* AND EXPLANATORY VARIABLE LIST.
* RUN MAXIMUM LIKELIHOOD PROCEDURE

ml model lf myprobit ( = w)
ml maximize
```

The line `args logl xb` is important. It indicates that the quantity that we wish to maximise is the sum over the sample of the variable named `logl`, and that the parameters with respect to which we wish to maximise it are implicit in the variable `xb`, which corresponds to $\beta_0 + \beta_1 w$ in the formula. `logl` and `xb` are examples of "local variables", being variables which exist within the program but not outside it. Any other local variables need to be declared by the `tempvar` command. Whenever temporary variables are referred to within the program, they need to be placed inside a particular set of quotation marks:

```
`p'
```

The quote before the p is the left single quote; you will find it in the upper left corner of most keyboards, below the "escape" key. The quote after the p is the right single quote; you will find it somewhere near the "enter" key.

Variables appearing without quotes are "global" variables, meaning that they also exist outside the program. In this example, yy (the binary dependent variable) is a global variable.

The last two lines of the above code are the lines that cause the program to run. The `ml` command specifies that the `lf` likelihood evaluator will be used. `lf` stands for "linear form", which essentially means that the likelihood evaluation program returns one log-likelihood contribution for each row of the data set. A situation in which the linear form restriction is not met is in the context of a panel data model, for which the likelihood evaluation program will return one contribution for each *block* of rows. In such a situation the *d-family* evaluators are required in place of `lf`. These will be introduced later.

The results from running the above code are:

```
. ml model lf myprobit ( = w)
. ml maximize

initial:       log likelihood = -727.80454
alternative:   log likelihood =  -635.1321
rescale:       log likelihood =  -635.1321
Iteration 0:   log likelihood =  -635.1321
Iteration 1:   log likelihood = -584.84039
Iteration 2:   log likelihood = -584.58503
Iteration 3:   log likelihood = -584.58503

                                                Number of obs   =       1050
```

```
                                                  Wald chi2(1)    =      94.05
Log likelihood = -584.58503                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |  -.1409882   .0145377    -9.70   0.000    -.1694816   -.1124948
       _cons |   1.301654   .0911155    14.29   0.000     1.123071    1.480237
------------------------------------------------------------------------------
```

Note that the results are identical to the results obtained using the command `probit y w` presented in Section 3.1.

See Exercise 1 for more practice with the `ml` routine.

3.4 Structural Modelling

In the present context, a structural model is one that is expressed in terms of the individual's utility function. The models analysed in earlier sections were not structural; they simply attempted to explain the data. Here, and in the next section, we consider the estimation of some simple structural models.

Firstly, let us assume that all individuals have the same utility function, which is:

$$\begin{aligned} U(x) &= \frac{x^{1-r}}{1-r} \qquad r \neq 1 \\ &= ln(x) \qquad r = 1 \end{aligned} \tag{33}$$

Equation (33) is known as the constant relative risk aversion (CRRA) utility function, because the parameter r is the coefficient of relative risk aversion: the higher is r, the more risk averse the subject. Note that r can be negative, indicating risk-seeking.

Secondly, let us assume that individuals maximise expected utility. We shall continue to use the "house money effect" example introduced in Section 3.1. The expected utilities from choosing the safe and risky choices are:

$$EU(S) = \frac{(w+5)^{1-r}}{1-r} \tag{34}$$

$$EU(R) = 0.5\frac{(w)^{1-r}}{1-r} + 0.5\frac{(w+10)^{1-r}}{1-r} \tag{35}$$

Thirdly, let us assume that when an individual computes the expected utility difference, they make a computational error ϵ, where $\epsilon \sim N(0, \sigma^2)$. This type of error has come to be known in risk modelling as a *Fechner error term* after Fechner (1860).

Given these three assumptions, the safe choice is made if:

$$EU(S) - EU(R) + \epsilon > 0$$

where $EU(S)$ and $EU(R)$ are defined in (34) and (35) respectively.

The probability of the safe choice being made is therefore:

$$\begin{aligned} P(S) &= P[EU(S) - EU(R) + \epsilon > 0] \\ &= P[\epsilon > EU(R) - EU(S)] \\ &= P\left[\frac{\epsilon}{\sigma} > \frac{EU(R) - EU(S)}{\sigma}\right] \\ &= 1 - \Phi\left[\frac{EU(R) - EU(S)}{\sigma}\right] \\ &= \Phi\left[\frac{EU(S) - EU(R)}{\sigma}\right] \end{aligned} \tag{36}$$

Substituting (34) and (35) into (36), and using the "yy trick" introduced in Section 3.1, the log-likelihood function may be written:

$$LogL = \sum_{i=1}^{n} ln\Phi\left[yy_i \times \frac{\frac{(w_i+5)^{1-r}}{1-r} - \left(0.5\frac{(w_i)^{1-r}}{1-r} + 0.5\frac{(w_i+10)^{1-r}}{1-r}\right)}{\sigma}\right] \tag{37}$$

We maximise (37) to obtain estimates of the two parameters r and σ. The challenge is that there is no STATA command that does this for us. We need to program it and use the `ml` command.

The required program, and the commands required to run the program, are as follows. For infomation about the syntax, the reader should refer back to the example provided in Section 3.3 in which each step was explained.

```
program drop structural
program structural
args logl r sig
tempvar eus eur diff p

quietly gen double `eus'=(w+5)^(1-`r')/(1-`r')
quietly gen double `eur'=0.5*w^(1-`r')/(1-`r')+0.5*(w+10)^(1-`r')/(1-`r')
quietly gen double `diff'=(`eus'-`eur')/`sig'
quietly gen double `p'=normal(yy*`diff')
quietly replace `logl'=ln(`p')
end

ml model lf structural /r  /sig
ml maximize
```

The line `args logl r sig` is again important. Here, it indicates that the quantity we are seeking to maximise is named `logl`, and that the parameters with respect to which we wish to maximise it are `r` and `sig`. One difference from the code in Section 3.3 is that the two parameters (`r` and `sig`) are named in the `ml` command. This is appropriate because these two parameters are stand-alone parameters, unlike those in the example in Section 3.3 which were regression parameters. Providing parameter names in the `ml` command is useful because it causes the same names to be included in the results table.

The results are as follows:

```
. ml model lf structural /r  /sig
```

```
. ml maximize

initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -601.45646
rescale:       log likelihood = -601.45646
rescale eq:    log likelihood = -600.78259
Iteration 0:   log likelihood = -600.78259
Iteration 1:   log likelihood =  -595.2424
Iteration 2:   log likelihood = -595.22797
Iteration 3:   log likelihood = -595.22739
Iteration 4:   log likelihood = -595.22739

                                                Number of obs   =       1050
                                                Wald chi2(0)    =          .
Log likelihood = -595.22739                     Prob > chi2     =          .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r            |
       _cons |     .21765   .0976928     2.23   0.026     .0261757    .4091244
-------------+----------------------------------------------------------------
sig          |
       _cons |   .3585733   .1046733     3.43   0.001     .1534174    .5637292
------------------------------------------------------------------------------
```

We see that the following estimates are obtained for the two parameters:

$$\begin{aligned} \hat{r} &= 0.2177 \\ \hat{\sigma} &= 0.3586 \end{aligned}$$

So, on the basis of the assumptions of this model, it appears that every individual is operating with the same utility function:

$$U(x) = \frac{x^{1-0.2177}}{1-0.2177} = \frac{x^{0.7823}}{0.7823}$$

and also that, when individuals compute the difference between the expected utilities of the two lotteries, they make a random computational error with mean zero and standard deviation 0.3586.

Of course, this is a homogeneous model assuming everyone having the same risk preferences. This assumption is relaxed in the next section.

3.5 Further Structural Modelling

3.6 The heterogeneous agent model

We continue to assume that subjects have the CRRA utility function:

$$U(x) = \frac{x^{1-r}}{1-r} \qquad r \neq 1$$

In Section 3.4, we assumed that all individuals had the same risk attitude, i.e. all had the same value of r. We attributed variation in choices to errors in the computation of expected utilities.

Here, we shall adopt a different approach. We shall assume (more realistically) that each subject has his or her own value of r, and we shall refer to the model as the "heterogeneous agent model". We just need to make an assumption about how r varies over the population. An obvious choice is:

$$r \sim N(\mu, \sigma^2) \tag{38}$$

We ask each subject to make a choice between two lotteries, S and R. We shall use the popular Holt & Laury (2002) design, which is presented in Table 1.

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

Table 1: The Holt and Laury design, with threshold risk aversion parameter for each choice problem

In Table 1, there are ten problems listed in order. In Problem 1, we expect all subjects to choose S; in Problem 10, we expect all subjects to choose R (in fact, R stochastically dominates in Problem 10). What is interesting is *where* in the sequence a subject switches from S to R, since this will indicate their attitude to risk. The content of Table 1 is sometimes called a "multiple price list" (MPL).

In the fourth column of Table 1, a value r^* is shown. This is known as the "threshold risk attitude" for the problem. It is the risk attitude (i.e. the coefficient of relative risk aversion) that would (assuming EU) make a subject indifferent between S and R for the choice problem. It can be worked out using Excel (see the spreadsheet: **risk aversion calculations**) as shown below.

	A	B	C	D	E	F	G	H	I	J
1	r:	-1.72	-0.95	-0.49	-0.15	0.15	0.41	0.68	0.97	1.37
2	prob of higher outcome:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3										
4	x	U(x)	U(x)	U(x)	U(x)	U(x)	U(x)	U(x)	U(x)	U(x)
5	0.1	0.000701	0.005754	0.021718	0.061561	0.166181	0.43566	1.495719	31.10848	-6.33575
6	1.6	1.320193	1.282329	1.351925	1.492932	1.754216	2.236551	3.632189	33.80667	-2.2713
7	2	2.422327	1.981408	1.885161	1.929686	2.120589	2.551266	3.901033	34.03374	-2.0913
8	3.85	14.38415	7.105814	5.002075	4.098095	3.700179	3.754653	4.81058	34.70904	-1.64126
9										
10	eu(S):	1.430406	1.422145	1.511896	1.667634	1.937403	2.42538	3.82038	33.98832	-2.1093
11	eu(R):	1.439046	1.425766	1.515825	1.676174	1.93318	2.427056	3.816122	33.98893	-2.11071
12										
13	cert equiv (S):	1.647867	1.687207	1.724713	1.761618	1.798329	1.835619	1.873599	1.912933	1.954209
14	cert equiv (R):	1.651519	1.689409	1.72772	1.76946	1.793719	1.83777	1.867081	1.914063	1.950689
15										

As an example, if a subject chooses S on problems 1–6, and chooses R on problems 7–10, they are revealing that (assuming EU) their risk attitude (r) is somewhere between 0.41 and 0.68.

Here, we assume that each subject is only asked to solve one of the ten problems. Each problem is solved by ten subjects, so we have 100 subjects in total. The data

is contained in the file **holtlaury_sim**.

Assume that subject i is presented with a choice problem with threshold risk level r_i^*. Let $y_i = 1$ if S is chosen, and $y_i = 0$ if R is chosen. The probability of subject i choosing S is (using the normal distribution of r specified in (38)):

$$\begin{aligned} P(y_i = 1) &= P(r_i > r_i^*) = P\left(z > \frac{r_i^* - \mu}{\sigma}\right) = P\left(z < \frac{\mu - r_i^*}{\sigma}\right) \\ &= \Phi\left(\frac{\mu - r_i^*}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma} - \left(\frac{1}{\sigma}\right) r_i^*\right) \qquad i = 1, \ldots, n \end{aligned} \tag{39}$$

In (39) we again have a probit model with dependent variable y. The explanatory variable is the threshold risk attitude for the problem being solved, r^*.

The intercept is $\frac{\mu}{\sigma}$ and the slope is $-\frac{1}{\sigma}$. Therefore from the probit estimates we are able to deduce estimates of μ and σ. This is done in STATA using the delta method (see next sub-section).

The output from the probit model is as follows:

```
. probit y rstar

Iteration 0:   log likelihood = -68.994376
Iteration 1:   log likelihood = -32.754689
Iteration 2:   log likelihood = -31.899974
Iteration 3:   log likelihood = -31.896643
Iteration 4:   log likelihood = -31.896643

Probit regression                                 Number of obs   =        100
                                                  LR chi2(1)      =      74.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -31.896643                       Pseudo R2       =     0.5377

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rstar |  -1.826082   .3481266    -5.25   0.000    -2.508398   -1.143767
       _cons |   .7306556   .2264169     3.23   0.001     .2868867    1.174424
------------------------------------------------------------------------------
Note: 10 failures and 0 successes completely determined.

.
. nlcom (mu: -_b[_cons]/_b[rstar]) (sig: -1/_b[rstar])

          mu:  -_b[_cons]/_b[rstar]
         sig:  -1/_b[rstar]

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          mu |    .400122   .0978294     4.09   0.000     .2083799    .5918641
         sig |   .5476205    .104399     5.25   0.000     .3430021    .7522389
------------------------------------------------------------------------------
```

Note that we have estimates of μ and σ. But this time, as we have estimated a heterogeneous agent model, the interpretation is as follows: every individual has a different "coefficient of relative risk aversion", drawn from the following distribution:

$$r \sim N\left(0.4001, 0.5476^2\right)$$

Having drawn their risk aversion parameter, they use this in the expected utility

calculation, which they perform without error.

3.7 The delta method

The delta method (`nlcom` in STATA) is used to obtain *standard errors* of the estimates of μ and σ in (39).

Let the probit estimates be $\hat{\beta}$ and $\hat{\alpha}$. We might refer to these estimates as the *reduced form* estimates. Using STATA, we can obtain an estimate of the variance matrix of these estimates:

$$\hat{V}\begin{pmatrix}\hat{\beta}\\ \hat{\alpha}\end{pmatrix}=\begin{pmatrix} var\left(\hat{\beta}\right) & cov\left(\hat{\alpha},\hat{\beta}\right)\\ cov\left(\hat{\alpha},\hat{\beta}\right) & var\left(\hat{\alpha}\right)\end{pmatrix} \tag{40}$$

The square roots of the diagonal elements of this matrix are the standard errors that we see in the STATA output from the `probit` command.

If you wanted to see $\hat{V}$ having estimated the probit model, you would do it as follows, and this is what you would see:

```
. mat V=e(V)
. mat list V

symmetric V[2,2]
                 y:          y:
              rstar       _cons
y:rstar    .12119211
y:_cons   -.04842685    .05126459
```

The parameters that we are interested in are functions of α and β.

$$\alpha=\frac{\mu}{\sigma};\quad \beta=-\frac{1}{\sigma}\Rightarrow\mu=-\frac{\alpha}{\beta};\quad \sigma=-\frac{1}{\beta} \tag{41}$$

We would refer to μ and σ as the *structural parameters*, being parameters of the utility function underlying behaviour.

We require the matrix D, where:

$$D=\begin{pmatrix}\frac{\partial\mu}{\partial\beta} & \frac{\partial\mu}{\partial\alpha}\\ \frac{\partial\sigma}{\partial\beta} & \frac{\partial\sigma}{\partial\alpha}\end{pmatrix}=\begin{pmatrix}\frac{\alpha}{\beta^2} & -\frac{1}{\beta}\\ \frac{1}{\beta^2} & 0\end{pmatrix} \tag{42}$$

Let $\hat{D}$ be the matrix D with parameters replaced by MLE's. The variance matrix of $\hat{\mu}$ and $\hat{\sigma}$ is:

$$\hat{V}\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix}=\hat{D}\left[\hat{V}\begin{pmatrix}\hat{\beta}\\ \hat{\alpha}\end{pmatrix}\right]\hat{D}' \tag{43}$$

The required standard errors are the square roots of the diagonal elements of this matrix.

Note that the delta method is applied using the `nlcom` command in STATA. This command will be used again. The example presented at the end of Section 3.6 makes clear the required syntax of the command.

3.8 Other Data Types

3.9 Interval data: the interval regression model

Let us return to the Holt & Laury (2002) design (Table 1). We continue to assume that subjects have the CRRA utility function:

$$U(x) = \frac{x^{1-r}}{1-r} \qquad r \neq 1$$

Recall that in the fourth column of Table 1, we show the value of r (the coefficient of relative risk aversion) that would make a subject indifferent between the two lotteries. Recall also that in Problem 1 we expect all subjects to choose S; in Problem 10, we expect all subjects to choose R. In Section 3.6, we considered ways of estimating the distribution of r over the population, when the available data consists of choices between pairs of lotteries.

In this section, we assume that the available information is more precise. We ask each subject to solve each choice problem in order, starting with Problem 1, thus revealing where in the list they switch from S to R. Under the assumption of EU, knowledge of where a subject switches gives us an interval for r for that subject. For example, an EU-maximising subject switching between Problems 5 and 6 is revealing that their coefficient of relative risk aversion is *between 0.15 and 0.41.*

The sort of data that results is known as "interval data". We are interested in the appropriate method for estimating the distribution of r over the population when interval data is available.

The file **interval_data_sim** contains this information for 100 subjects (as well as information on subject characteristics).

As in Section 3.4, we assume that the distribution of r over the population is:

$$r \sim N(\mu, \sigma^2) \tag{44}$$

For each subject, i, we have a lower bound (l_i) and an upper bound (u_i) for his or her r-value. The likelihood contribution for each subject is the probability of them being in the interval in which they are observed. So:

$$L_i = P(l_i < r < u_i) = P(r < u_i) - P(r < l_i) = \Phi\left(\frac{u_i - \mu}{\sigma}\right) - \Phi\left(\frac{l_i - \mu}{\sigma}\right) \tag{45}$$

So the sample log-likelihood is:

$$LogL = \sum_{i=1}^{n} \left[\Phi\left(\frac{u_i - \mu}{\sigma}\right) - \Phi\left(\frac{l_i - \mu}{\sigma}\right) \right] \tag{46}$$

Equation (46) is maximised to give MLEs of μ and σ.
This is called the interval regression model (although at present there are no explanatory variables). To estimate it in STATA, use the command:

```
intreg  rlower rupper
```

where `rlower` and `rupper` are the variables containing the lower and upper bounds for each observation.
The results are as follows:

```
. intreg rlower rupper

Fitting constant-only model:

Iteration 0:   log likelihood = -199.07231
Iteration 1:   log likelihood = -198.96851
Iteration 2:   log likelihood = -198.96849

Fitting full model:

Iteration 0:   log likelihood = -198.96849
Iteration 1:   log likelihood = -198.96849

Interval regression                             Number of obs   =        100
                                                LR chi2(0)      =       0.00
Log likelihood = -198.96849                     Prob > chi2     =          .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    .613146   .0597808    10.26   0.000     .4959777    .7303143
-------------+----------------------------------------------------------------
    /lnsigma |  -.5323404   .0764651    -6.96   0.000    -.6822092   -.3824716
-------------+----------------------------------------------------------------
       sigma |    .587229   .0449025                      .505499    .6821733
------------------------------------------------------------------------------

  Observation summary:         0  left-censored observations
                               0     uncensored observations
                               6 right-censored observations
                              94       interval observations
```

Estimates of the parameters of interest can be read directly. The distribution of risk-attitude over the population is estimated to be:

$$r \sim N\left(0.613, 0.587^2\right)$$

Next, suppose that we wish to allow risk attitude to vary according to subject characteristics. For example:

$$\begin{aligned} r_i &= \beta_0 + \beta_1 age_i + \beta_2 male_i + \epsilon_i \\ &= x_i'\beta + \epsilon_i \\ \epsilon_i &\sim N\left(0, \sigma^2\right) \end{aligned} \tag{47}$$

In the second line of (47) we are adopting the convention of collecting all explanatory variables pertaining to observation i, including a constant, into the vector x_i. The vector β contains the three corresponding parameters $\beta = (\beta_0 \quad \beta_1 \quad \beta_2)'$.
With this generalisation, (44) becomes:

$$r_i \sim N\left(x_i'\beta, \sigma^2\right) \tag{48}$$

and the log-likelihood function becomes:

$$LogL = \sum_{i=1}^{n} ln \left[\Phi \left(\frac{u_i - x_i'\beta}{\sigma} \right) - \Phi \left(\frac{l_i - x_i'\beta}{\sigma} \right) \right] \tag{49}$$

To estimate an interval regression model with explanatory variables, we do as follows:

```
. intreg rlower rupper age male

Fitting constant-only model:

Iteration 0:   log likelihood = -199.07231
Iteration 1:   log likelihood = -198.96851
Iteration 2:   log likelihood = -198.96849

Fitting full model:

Iteration 0:   log likelihood = -197.24143
Iteration 1:   log likelihood = -197.17109
Iteration 2:   log likelihood = -197.17108

Interval regression                             Number of obs   =        100
                                                LR chi2(2)      =       3.59
Log likelihood = -197.17108                     Prob > chi2     =     0.1657

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |     .02213   .0196956     1.12   0.261    -.0164727    .0607327
        male |  -.2165679   .1341118    -1.61   0.106    -.4794222    .0462864
       _cons |   .1592841   .4565128     0.35   0.727    -.7354646    1.054033
-------------+----------------------------------------------------------------
    /lnsigma |  -.5507208   .0764747    -7.20   0.000    -.7006085   -.4008332
-------------+----------------------------------------------------------------
       sigma |   .5765341   .0440903                      .4962832    .6697618
------------------------------------------------------------------------------

  Observation summary:         0  left-censored observations
                               0     uncensored observations
                               6 right-censored observations
                              94       interval observations
```

From these results, we now have an equation which determines the risk-aversion parameter for an individual of a given age and gender:

$$\hat{r}_i = 0.159 + 0.022 age_i - 0.217 male_i$$

However, note that neither of these explanatory variables appear to have significant effects on risk attitude. "Male" is close to being significant, and the negative sign would tell us that males are less risk-averse (or more risk-seeking) than females.

3.10 Continuous (exact) data

Yet another way of eliciting risk attitude is to present a subject with a single lottery, and ask them for their "certainty equivalent", that is, the amount of money such that they would be exactly indifferent between receiving this sum of money and playing the lottery.

For example, if the lottery is:

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

and the subject claims that their certainty equivalent is \$0.75, then we can deduce that their coefficient of relative risk aversion is exactly 0.41. How do we know this? Because:[8]

$$0.3\frac{3.85^{1-0.41}}{1-0.41}+0.7\frac{0.10^{1-0.41}}{1-0.41}=\frac{0.75^{1-0.41}}{1-0.41}$$

A very important question is: how do you elicit a subject's certainty equivalent? You can simply ask them, and hope that they give an honest answer. But, according to some, there needs to be an *incentive* for the subject to report their certainty equivalent correctly. The scheme used to elicit the certainty equivalent needs to be *incentive compatible*.

One popular method for doing this, which is under reasonable assumptions incentive compatible, is the Becker-DeGroot-Marschak (BDM; Becker et al. 1964) incentive mechanism. BDM is described as follows. The individual is asked to place a valuation on a lottery (i.e. to report their certainty equivalent). They are told that after they have done this a random "price" will be generated. If the randomly generated price is higher than their reported valuation, they will be given an amount of money equal to this price, and they will not play the lottery; if the price is lower than their valuation, they will play the lottery.

In the file **exact_data_sim**, we have the values of r elicited in this way for 100 subjects. This is "exact" data in the sense that the value of r is exactly observed. It is also "continuous" data as opposed to "discrete" (binary and interval data are both forms of "discrete" data). The distribution of r over the sample of 100 subjects is shown in Figure 10.

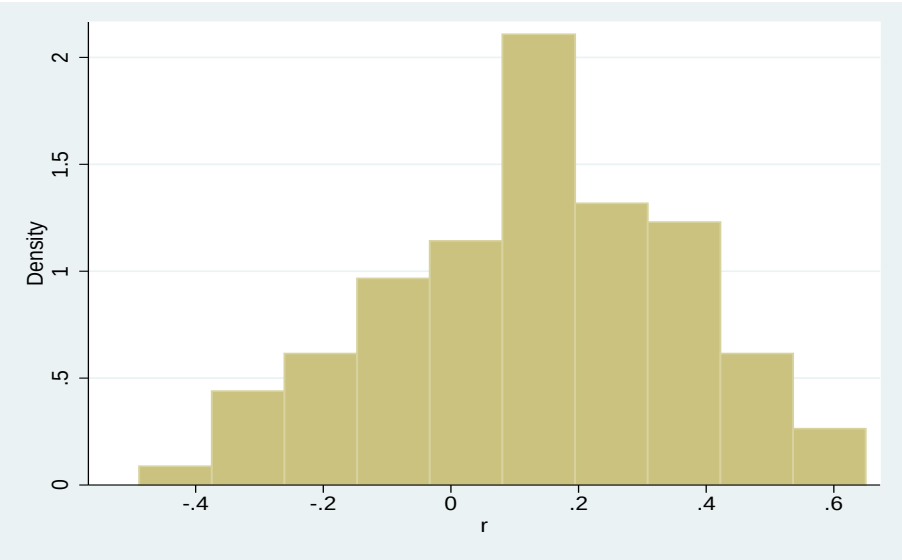


Figure 10: The distribution of r over 100 subjects

We return to the assumption:

$$r\sim N\left(\mu,\sigma^2\right) \tag{50}$$

How do we estimate μ and σ when exact data is available? First, let us consider

[8]This may be verfied easily using the Excel sheet "risk aversion calculations".

what happens when we try to do this using maximum likelihood.

Consider the density associated with a particular observation r_i:

$$f(r_i;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(r_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{r_i-\mu}{\sigma}\right) \tag{51}$$

Equation (51) is a typical likelihood contribution, so the sample log-likelihood function is given by:

$$LogL = \sum_{i=1}^{n} ln\left[\frac{1}{\sigma}\phi\left(\frac{r_i-\mu}{\sigma}\right)\right] \tag{52}$$

To program (52), we would do as follows:

```
program define exact
args lnf xb sig
tempvar y p

quietly gen double ‘y’=$ML_y1
quietly gen double ‘p’=(1/‘sig’)*normalden((‘y’-‘xb’)/‘sig’)
quietly replace ‘lnf’=ln(‘p’)
end

ml model lf exact (r= ) ()
ml maximize
```

The results are as follows:

```
. ml maximize

initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -60.251905
rescale:       log likelihood = -7.5739988
rescale eq:    log likelihood =  3.1167494
Iteration 0:   log likelihood =  3.1167494
Iteration 1:   log likelihood =  3.2682025
Iteration 2:   log likelihood =  3.6372157
Iteration 3:   log likelihood =   3.637384
Iteration 4:   log likelihood =   3.637384

                                                Number of obs   =        100
                                                Wald chi2(0)    =          .
Log likelihood =   3.637384                     Prob > chi2     =          .

------------------------------------------------------------------------------
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
       _cons |   .1340463   .0233327     5.74   0.000     .0883149    .1797776
-------------+----------------------------------------------------------------
eq2          |
       _cons |   .2333275   .0164987    14.14   0.000     .2009905    .2656644
------------------------------------------------------------------------------
```

and we see that the maximum likelihood estimates are:

$$\hat{\mu} = 0.134$$
$$\hat{\sigma} = 0.233$$

Of course, there is a much easier way of obtaining the MLEs of μ and σ when exact data is available:

```
. summ r
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           r |       100    .1340463    .2345029  -.4884877   .6499107
```

The maximum likelihood estimates of μ and σ are just the sample mean and sample standard deviation respectively, of the variable r. The slight difference between the MLE of σ and the sample standard deviation arises because the former uses n as the divisor, while the latter uses $n-1$. Asymptotically, the two will be equal.

Note that the main purpose of using ML above is to remind ourselves of the structure of the log-likelihood function in a situation in which the data are *continuous*. This is particularly important in the next sub-section, in which we consider *censored data*, which usually takes the form of a mixture of discrete and continuous data.

What is interesting about the results is that the estimate of μ (0.134) is much closer to zero than the estimates obtained using all other methods so far (which have varied between 0.400 and 0.613). This suggests that, while we know that most subjects are risk averse when choosing between lotteries, they tend towards risk neutrality when asked for certainty equivalents (when r=0, we would have risk neutrality). Another way of saying this is that when a subject is asked for a certainty equivalent, there is a tendency for them to compute the *expected value* of the lottery, and to report something close to this.

The tendency towards risk-neutrality in valuation problems is an obvious explanation for the well-known "preference reversal" phenomenon discussed in Section **??**. This is the tendency for subjects to prefer the safer lottery (the "P-bet") when asked to choose between them, but to place a higher valuation on the riskier lottery (the "$-bet").

3.11 Further Analysis of Ultimatum Game Data

3.11.1 Tests of gender effects

We will return to the data on the ultimatum game, **ug_sim**.

It is possible to use treatment tests to test for the effect of gender. Just treat gender as the "treatment". However, a more informative way of looking for a gender effect is using regression analysis. The important advantage of regression analysis is that it enables us to estimate different effects at the same time.

For example, we might wish to do the following. We start by generating a dummy variable indicating whether a male proposer is giving to a female responder.

```
. gen m_to_f=male_i*(1-male_j)

.
. regress y male_i male_j m_to_f

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =    3.37
       Model |  976.185392     3  325.395131            Prob > F      =  0.0195
    Residual |  18901.4946   196   96.436197           R-squared     =  0.0491
-------------+------------------------------           Adj R-squared =  0.0346
```

```
       Total |    19877.68   199  99.8878392           Root MSE      =  9.8202

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      male_i |  -4.519608   1.885099    -2.40   0.017     -8.23729   -.8019261
      male_j |   3.744608   2.074081     1.81   0.073    -.3457722    7.834988
      m_to_f |   2.381863    2.80275     0.85   0.396    -3.145557    7.909282
       _cons |     35.275   1.552709    22.72   0.000     32.21284    38.33716
------------------------------------------------------------------------------
```

These results tell us that:

1. Male proposers tend to offer 4.5 units less than female proposers, *ceteris paribus*.

2. Proposers tend to offer 3.7 units more when the responder is male, than when the responder is female, *ceteris paribus*. Note that this effect is only marginally significant.

3. Male proposers tend to offer 2.38 units more when the responder is female than when the responder is male, *ceteris paribus*. Eckel & Grossman (2001) refer to this effect as the"chivalry effect". Note that the effect is not statistically significant in this sample.

In consideration of conclusion 2, that proposers offer more when the responder is male, we might ask whether it is rational to do so. It is rational to offer more to male responders if males are more likely to reject offers. To see if this is the case, we go back to the probit model of Section 3.2, and add gender of responder as an explanatory variable in addition to proposer's offer. The results from doing so are:

```
. probit d y male_j

Iteration 0:   log likelihood = -113.55237
Iteration 1:   log likelihood = -68.373743
Iteration 2:   log likelihood = -64.187937
Iteration 3:   log likelihood = -64.116934
Iteration 4:   log likelihood = -64.116904
Iteration 5:   log likelihood = -64.116904

Probit regression                                 Number of obs   =        200
                                                  LR chi2(2)      =      98.87
                                                  Prob > chi2     =     0.0000
Log likelihood = -64.116904                       Pseudo R2       =     0.4354

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |   .1567836   .0231961     6.76   0.000       .11132    .2022472
      male_j |  -.5976406   .2668131    -2.24   0.025    -1.120585   -.0746966
       _cons |  -3.933341   .6589175    -5.97   0.000    -5.224796   -2.641886
------------------------------------------------------------------------------
```

We see that there is indeed evidence ($p = 0.025$) that a male is less likely to accept an offer of a given size than a female. So we may conclude that it is rational for the proposer to offer more to male responders.

The obvious follow-up question is: how much more should a proposer offer to a male responder than to a female responder, in order to create the same propensity for the offer to be accepted? We see that male responders' propensity to accept is lower by 0.598. This difference may be restored by increasing the offer by an amount 0.598/0.157. This computation may of course be carried out using the `nlcom` command:

```
. nlcom more_to_male: -_b[male_j]/_b[y]

more_to_male:  -_b[male_j]/_b[y]

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
more_to_male |   3.811882   1.612015     2.36   0.018      .6523915    6.971373
------------------------------------------------------------------------------
```

We see that a rational proposer should offer 3.81 more to a male than to a female, although the 95% confidence interval for this quantity is fairly wide.

It is interesting that this estimate of 3.81 for the additional amount that a rational proposer should give to a male in order to restore the acceptance probability, is very close to the estimate of 3.74 obtained from the above regress command, of the additional amount that proposers actually do offer to males. It appears that proposers in this (simulated) data are indeed rational.

3.11.2 The proposer's decision as a risky choice problem

Let us remind ourselves of the relationship between the proposer's offer (y) and the responder's decision on whether to accept the offer (d). In Figure 9 we demonstrated that the probability of acceptance rises steeply with the amount of the offer, apparently reaching one when the offer reaches 50% of the endowment (recall that the total endowment was 100 tokens).

So, we could view the proposer's decision like this. If the proposer offers 50, he will keep 50 for himself with probability one; he will have a risk-free pay-off of 50. If he offers only 40, his pay-off will rise to 60, but he will receive this with probability less than 1, and otherwise he will receive zero. The lower the offer, the higher the possible pay-off, but the lower the probability of receiving this pay-off. Hence we see that the proposer's decision can be analysed as a risky choice problem. This is the approach taken by Roth et al. (1991), and others.

In Section 3.2, the probit model was used to obtain the following formula for the probability of any offer y being accepted:

$$\hat{P}(d=1) = \Phi(-3.855 + 0.144y)$$

Let us assume that the proposer knows this probability formula. Note that this amounts to the assumption of rational expectations.

For the present purpose, let us treat the total endowment as one unit. So, if the proposer offers 50% of the endowment, their risk-free pay-off will be 0.5. If they offer 40%, their uncertain payoff will be 0.6, and so on.

The Excel sheet **proposer decision** contains the calculations necessary for the following analysis. If we assume a particular risk aversion parameter, say $r = 0.4$, then we have an expression for the proposer's expected utility from offering y:

$$EU(\mathrm{y}) = \Phi(-3.855 + 0.144y) \times \left(\frac{100 - y}{100}\right)^{1-0.4} \Big/ (1 - 0.4) \qquad (53)$$

Using (53), we can plot EU against each possible offer. This is done in Figure 11, from which we can see that the optimal offer for a proposer with $r = 0.4$ is 40. By repeating this exercise for different values of r, we can find the optimal offer for each risk attitude. The result is plotted in Figure 12.

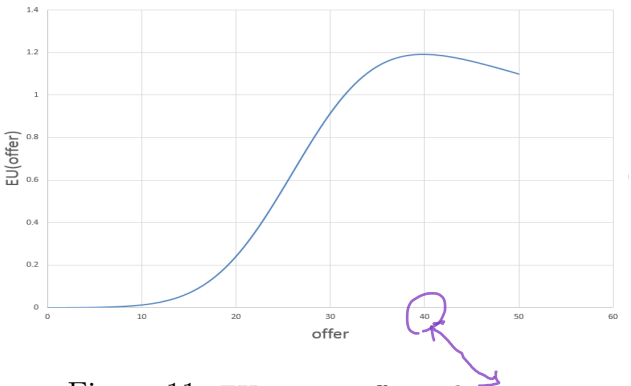


Figure 11: EU against offer with r=0.4

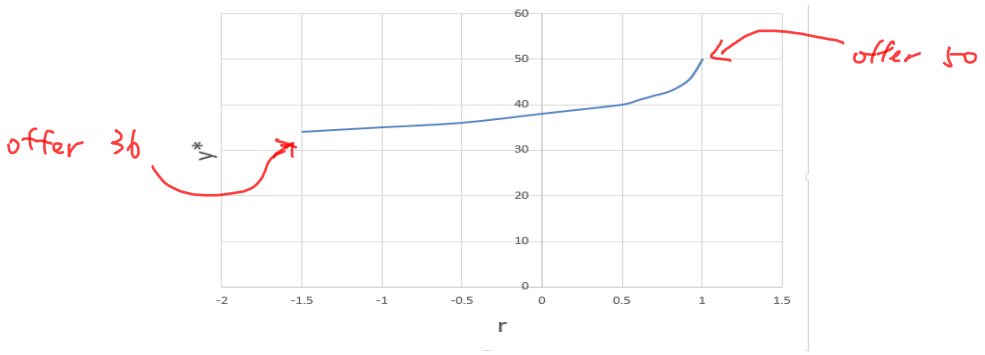


Figure 12: Optimal offer against r

We can then use Figure 12 to deduce a proposer's risk-aversion parameter from knowledge of their offer. For example, if they offer 47, their risk-aversion parameter must be 0.95 (they are quite risk averse).

Recall that 36 (18%) of the 200 proposers offered exactly 50% of their endowment. Should we attribute this behaviour to extreme risk aversion? Perhaps not. Individuals who offer 50% are likely to be doing so out of fairness considerations. They want to give 50% because they think it is the fair allocation, not because they are worried about having an offer rejected.

This sort of consideration leads us to a *mixture model*. Mixture models will be introduced in Part V. One group from the population (around 18% it seems) are motivated by fairness and wish to share the endowment equally. The other 82% are motivated by self-interest and their degree of risk aversion dictates how much they offer, in accordance with the analysis above.

Extending this idea, it is interesting to take a close look at the subjects who are motivated by fairness (let us label them "egalitarians"; they are sometimes also re-

ferred to as "equal-splitters").

```
. gen egal=y==50

. tab egal

      egal |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |        164       82.00       82.00
         1 |         36       18.00      100.00
-----------+-----------------------------------
     Total |        200      100.00
```

As remarked above, 36 proposers offer 50%. Let us now investigate how these 36 divide by gender. To investigate whether egalitarianism is related to gender, we require a chi-squared test (~~see Part I~~).

```
. tab male_i egal , chi2

           |         egal
    male_i |         0          1 |     Total
-----------+----------------------+----------
         0 |        66         25 |        91
         1 |        98         11 |       109
-----------+----------------------+----------
     Total |       164         36 |       200

          Pearson chi2(1) =  10.1506   Pr = 0.001
```

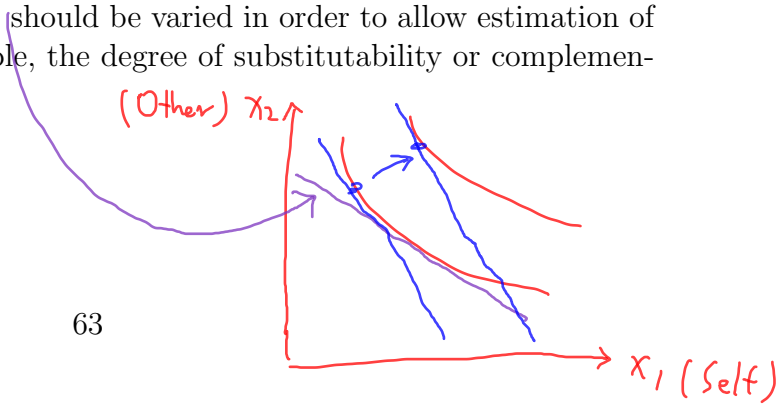
As we see in the table, 25 of the 91 females are egalitarian, while only 11 of the 109 males are so. The significance of this difference is summarised with the chi-squared test, and the accompanying p-value indicates a strongly significant relationship between gender and egalitarianism: females are significantly more likely than males to be egalitarian (according to this simulated data set).

4 Social Preference Models

4.1 Introduction

Here, we are concerned with the estimation of the parameters of the utility function underlying dictator game giving. That is, we focus on the *structure* of the preferences underlying behaviour. In this context, it has become conventional to assume that utility is an increasing function of two arguments: own payoff (x_1) and other's payoff (x_2).

In order to estimate the parameters of such a utility function, it is obviously essential for the experimental design to include variation in the endowment. However, it is also desirable for the prices of the two "goods" to vary. That is, the price of "giving" and the price of "keeping" should be varied in order to allow estimation of parameters representing, for example, the degree of substitutability or complementarity between these two "goods".



4.2 Estimation of Preference Parameters from Dictator Game Data

4.2.1 The framework

The experimental setting considered here is that of Andreoni & Miller (2002). In this setting, each individual is given an endowment (m) which they are required to allocate between "self" and "other", with both of these "goods" having a "price". For example, if "giving to self" has price $\frac{1}{2}$, the amount actually received by "self" will be twice the amount allocated; if "giving to other" has price $\frac{1}{3}$, the amount actually received by "other" will be three times the amount allocated.

We define the following variables:

x_1 = amount *received* by self
x_2 = amount *received* by other
m = endowment
p_1 = "price" of x_1 (i.e. for each unit of the endowment that you direct to yourself, you receive $1/p_1$ units).
p_2 = "price" of x_2 (i.e. for each unit of the endowment that you direct to the other player, they receive $1/p_2$ units).

An important point to stress at this stage is that, although x_1 and x_2 will be the two arguments of the dictator's utility function, they are *not* decision variables. The decision variables are, in fact:

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

Of course, these two decision variables are not both free variables. They are constrained by the budget constraint:

$$p_1x_1 + p_2x_2 \leq m \tag{54}$$

We will normally refer to the amount directed to other, p_2x_2 as the single decision variable, and recognise that, because the budget constraint will always be binding, the other decision variable is determined as $p_1x_1 = m - p_2x_2$.

It is also useful to define "budget shares": $w_1 = \frac{p_1x_1}{m}$; $\quad w_2 = \frac{p_2x_2}{m}$

4.2.2 The Andreoni-Miller data

Andreoni & Miller's (2002) data is contained in the file **garp**. The variables are as defined in Section 4.2.1. Here we shall provide a description of the data, and report on some exploratory analysis thereof.

There were 176 subjects in the experiment. Each subject was faced with a sequence of decision problems in the form of budgets. Each budget had a different combination of endowment (m), price of keeping (p_1), and price of giving (p_2). These combinations are shown in Table 2. The task subjects are required to perform for each budget is to decide how much of the endowment (m) to keep for themselves

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

Standard dictator game

Table 2: Andreoni & Miller's (2002) design.
Notes: There are 11 different budgets, each with a different combination of endowment (m), price of keeping (p_1), and price of giving (p_2). In budgets 1-8, all 176 subjects participated; in budgets 9-11, only 34 participated. Final column shows mean amount sent to other.

(p_1x_1), and how much to send to the other player (p_2x_2). The decision problems were presented in a random order to each subject. Subjects were told that, when all decisions had been made, one of the decision problems would be chosen at random and carried out with another randomly chosen subject as the recipient.

Budgets 1–8 were faced by all 176 subjects. Budgets 9–11 were only faced by 34 of the subjects. The final column of Table 2 shows the average amount sent to the other player for each budget. Clearly, there is a good amount of variation in this outcome.

Note from Table 2 that for three of the problems, 7, 8, and 9, the two prices are both one, implying that these tasks correspond to the standard dictator game. From the second and sixth columns, we see that in these three tasks, average giving is between 17–24% of the endowment, in close agreement with previous dictator game experiments (Camerer 2003).

Figure 13 shows a (jittered) scatter of amount received by other against amount received by self. This further highlights the wide variation in the amount of giving. This is partly a result of the richness of the design (with 11 different budget constraints), and also a high apparent variation in preference for giving. As expected, there is a higher concentration of points in the lower-right region of the plot, reflecting an overall bias towards giving-to-self (with 42% of the observations on giving-to-other being zero). The "jitter" option has been used for the scatter so that clusters of observations at particular points (e.g. on the horizontal axis) are easily discernible.

Figure 14 shows a scatter of amount directed to other against endowment, with smoother. The positive relationship evident here simply indicates that giving is a "normal good".

```
scatter x2 x1, msize(0..3) jitter(1)
```

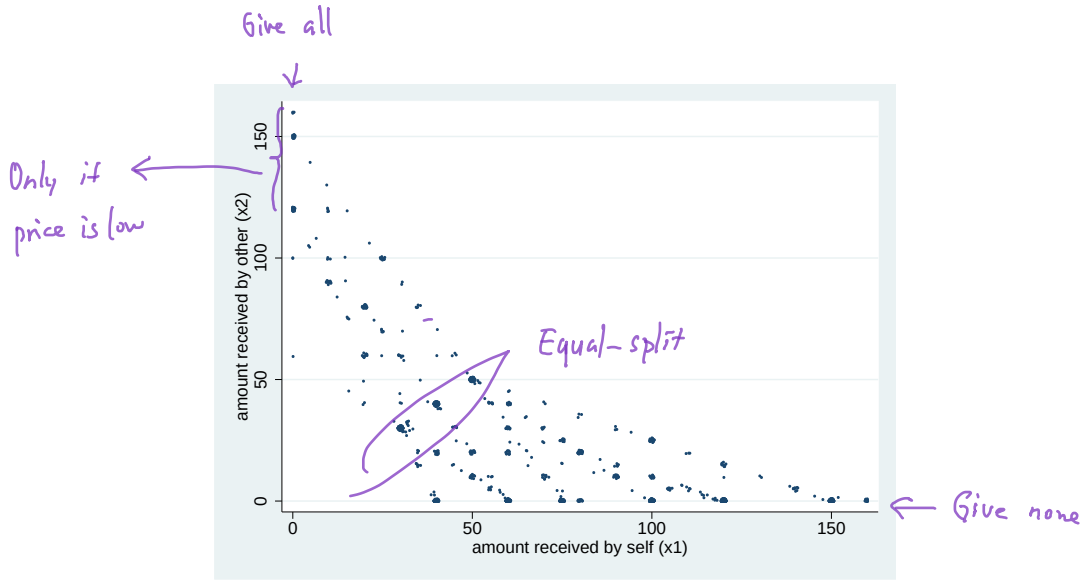


Figure 13: Jittered scatter of data in (x_1, x_2) space.

```
lowess p2x2 m
```

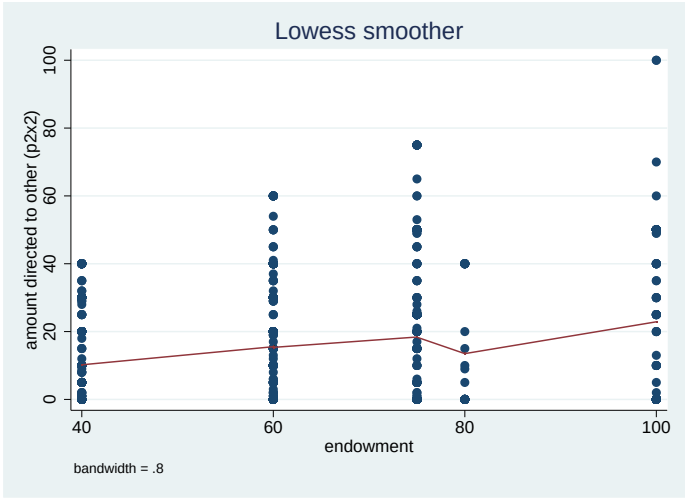


Figure 14: Amount directed to other against endowment

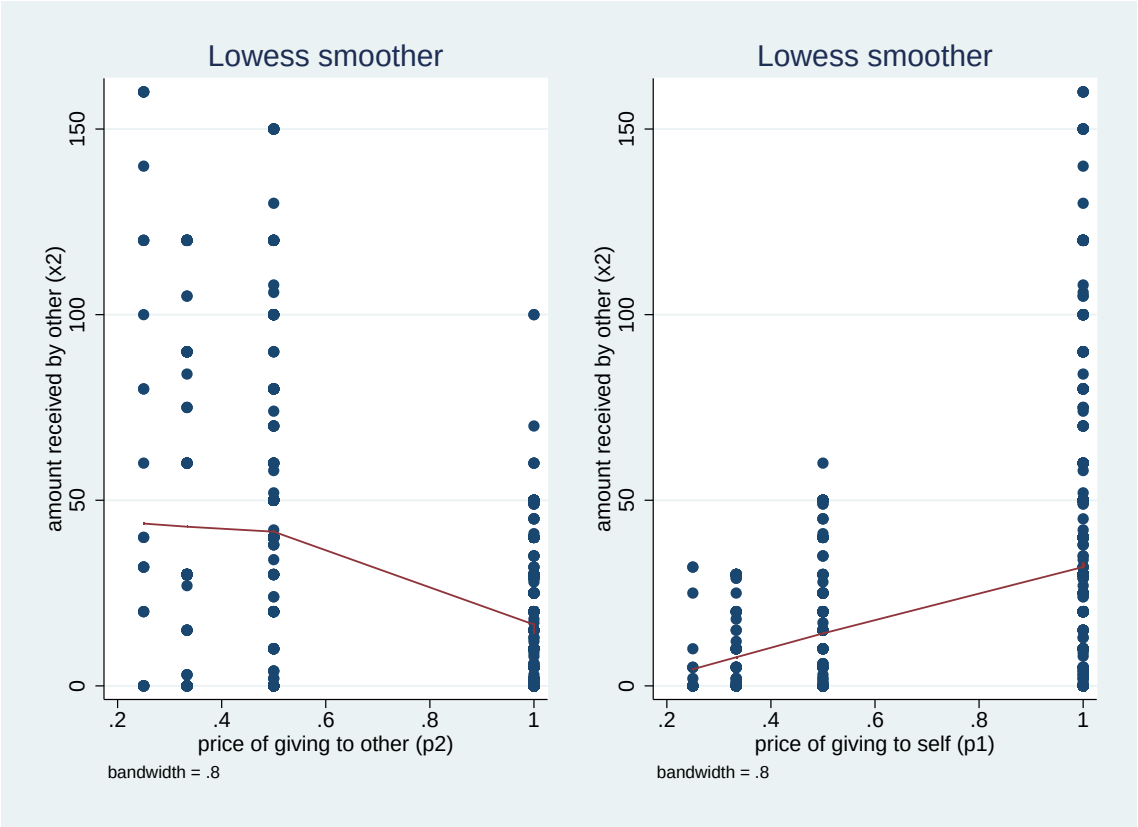


Figure 15: Amount received by other against price of giving to other (left pane), and against price of giving to self (right pane), with smoothers.

Figure 15 shows smoothers of amount *received* by other against price of giving to other (left pane) and against price of giving to self (right pane). The downward-sloping curve seen in the left-hand graph is consistent with the "law of demand"; the upward-sloping curve seen in the right-hand graph is consistent with the two "goods" (amount received by other and amount received by self) being substitutes.

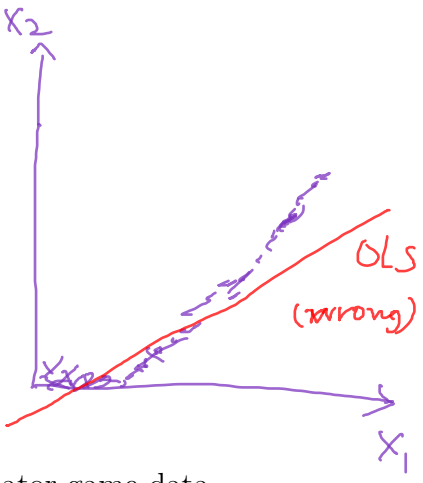
These results may be confirmed using a linear regression. Regression of amount received by other on the two prices, with clustering at the subject level, gives the following results. The effects of the two prices are very strong, and with signs as expected on the basis of the two graphs in Figure 15.

```
. regress x2 p2 p1, vce(cluster i)

Linear regression                               Number of obs =    1510
                                                F(  2,   175) =   61.20
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1847
                                                Root MSE      =  28.661

                                  (Std. Err. adjusted for 176 clusters in i)
------------------------------------------------------------------------------
             |               Robust
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p2 |  -39.00726   4.934956    -7.90   0.000    -48.74695   -29.26757
          p1 |   14.47704   1.664276     8.70   0.000      11.1924    17.76167
       _cons |   43.95138   4.663821     9.42   0.000     34.74681    53.15596
------------------------------------------------------------------------------
```

Next we add income (i.e. the endowment) into the regression. Income is seen to

have a strongly positive effect, confirming that giving to other is a "normal good". The interpretation of its coefficient (0.265) is that when the dictator's endowment rises by one unit, *ceteris paribus*, amount received by other will rise by around one quarter of one unit. However, a consequence of adding income to the regression is that the effect of the price of giving to "self" is no longer significant. This is partly a result of the strong positive correlation between m and p_1 essentially causing p_1 to take the role of "proxy" for m in the model in which the latter is excluded.

```
. regress x2 p2 p1 m, vce(cluster i)

Linear regression                               Number of obs =    1510
                                                F(  3,   175) =   61.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1976
                                                Root MSE      =  28.441

                                  (Std. Err. adjusted for 176 clusters in i)
------------------------------------------------------------------------------
             |               Robust
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p2 |  -52.12677   5.063235   -10.30   0.000    -62.11964   -42.13391
          p1 |   1.357528   1.783083     0.76   0.447    -2.161587    4.876643
           m |    .265248   .0277023     9.57   0.000     .2105744    .3199216
       _cons |   47.92717   4.707122    10.18   0.000     38.63713    57.2172
------------------------------------------------------------------------------
```



It is obvious from Figure 13, and also from previous analysis of dictator game data within this course, that there is an accumulation of zero observations in giving to other. Around 42% of this sample consist of observations with giving equal to zero. The linear regressions just performed do not take account of this accumulation of zero observations. A model which does take account of this feature of the data is the Tobit model, explained in detail in Section **??**. We next estimate a Tobit model of giving on the two prices and income, again with cluster-robust standard errors. The results are as follows.

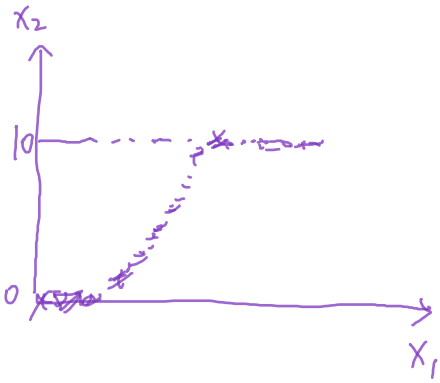
```
. tobit x2 p2 p1 m, vce(cluster i) ll(0)

Tobit regression                                Number of obs   =       1510
                                                F(  3,  1507)   =      54.33
                                                Prob > F        =     0.0000
Log pseudolikelihood =  -5027.146               Pseudo R2       =     0.0256

                                  (Std. Err. adjusted for 176 clusters in i)
------------------------------------------------------------------------------
             |               Robust
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p2 |  -67.1347   7.049639    -9.52   0.000    -80.96285   -53.30656
          p1 |   10.8052   3.910197     2.76   0.006     3.135191    18.4752
           m |   .3322818   .0380964     8.72   0.000     .2575541    .4070095
       _cons |   34.41715   6.122105     5.62   0.000      22.4084    46.4259
-------------+----------------------------------------------------------------
      /sigma |   42.59774    2.46888                      37.75494    47.44055
------------------------------------------------------------------------------
  Obs. summary:        628  left-censored observations at x2<=0
                       882     uncensored observations
                         0 right-censored observations
```

We see that the coefficient estimates from Tobit are considerably larger in magnitude than the corresponding OLS estimate. Most strikingly, the Tobit coefficient of price of amount received by self (p_1) is 10.81, eight times larger than the OLS estimate of the same parameter, which is 1.36. Furthermore, the Tobit coefficient is

strongly significant ($p = 0.006$) compared to a complete lack of significance under OLS ($p = 0.447$). This emphatically confirms the importance of dealing with zero censoring when analysing data sets of this type. This will be the focus in Section **??**.

Of course, we can go one step further and estimate the random effects Tobit model. The results are:

```
. xtset i t
 . xttobit x2 p2 p1 m,  ll(0)

Random-effects Tobit regression                 Number of obs      =      1510
Group variable: i                               Number of groups   =       176

Random effects u_i ~ Gaussian                   Obs per group: min =         8
                                                               avg =       8.6
                                                               max =        11

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(3)       =    605.11
Log likelihood  = -4663.2072                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p2 |  -75.14353   4.942489   -15.20   0.000    -84.83063   -65.45643
          p1 |   9.896787   5.060785     1.96   0.051    -.0221691    19.81574
           m |   .3672872   .0639333     5.74   0.000     .2419803    .4925941
       _cons |   32.68706   6.512942     5.02   0.000     19.92193    45.4522
-------------+----------------------------------------------------------------
    /sigma_u |    44.0585   3.276081    13.45   0.000      37.6375     50.4795
    /sigma_e |   28.67666   .7433699    38.58   0.000     27.21968    30.13364
-------------+----------------------------------------------------------------
         rho |   .7024244   .0320737                      .6367994    .7620325
------------------------------------------------------------------------------
```



The importance of between-subject heterogeneity is clearly seen from the large and significant estimate of σ_u, of 44.06. The estimates are different as well: some of the slope estimates, particularly that of price of giving to other, take even larger values as a result of allowing for the heterogeneity. $-75.14 > -67.13$

4.2.3 Estimating the parameters of a CES utility function

In this section we will use the data set introduced in the previous sub-section to estimate the parameters of a utility function for altruism.

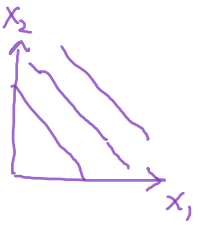
Following Andreoni & Miller (2002) and others, we will assume the constant elasticity of substitution (CES) utility function:

$$U(x_1, x_2) = [\alpha x_1^{\rho} + (1 - \alpha) x_2^{\rho})]^{\frac{1}{\rho}} \quad 0 \leq \alpha \leq 1 \quad -\infty \leq \rho \leq 1 \tag{55}$$

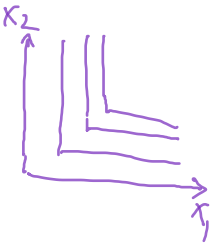
The CES utility function (55) is used in many areas of economics. In the present setting, the parameter α indicates selfishness, while the parameter ρ indicates willingness to trade off equity and efficiency in response to price changes. Values of ρ less than zero indicate a concern for equality in payoffs; values of ρ between zero and one indicate a focus on efficiency. The *elasticity of substitution*, usually denoted σ, may be deduced directly from ρ using:

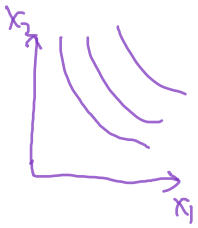
$$\sigma = \frac{1}{1 - \rho} \tag{56}$$

σ is clearly an increasing function of ρ, with values of ρ between zero and one (indicating a focus on efficiency) being associated with values of σ between one and $+\infty$.



A useful way of interpreting the elasticity of substitution, σ, is in terms of the curvature of the indifference curves. The larger is σ, the less curved the indifference curves become. As σ approaches $+\infty$ the indifference curves become downward-sloping straight lines, implying that the two goods are perfect substitutes, and that all that matters is the total payoff. At the other extreme, if σ appraoches its lower limit of zero, the indifference curves become L-shaped, implying perfect complements, and that all that matters is equality of payoffs. The intermediate case is when $\sigma = 1$, implying Cobb-Douglas preferences: $U = x_1^{\alpha} x_2^{1-\alpha}$.





Maximising (55) subject to the budget constraint (54), we arrive at the "Marshallian demand fuction" for own pay-off:

$$w_1 = \frac{p_1^{\frac{\rho}{\rho-1}}}{p_1^{\frac{\rho}{\rho-1}} + \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}} p_2^{\frac{\rho}{\rho-1}}} + \epsilon \tag{57}$$

where w_1 is, as previously mentioned, the share of the total allocation that is allocated to "self"; that is: $w_1 = \frac{p_1 x_1}{m}$. Note that a stochastic term (ϵ) has been appended in (57) in order to turn the deterministic budget-share equation into an estimable model. The equation for the second budget share w_2 could easily be deduced from the deterministic part of (57) because $w_2 = 1 - w_1$. However, we only need one of the two equations to estimate the two parameters. We will use (57).

Non-linear least squares is required to estimate the two parameters in (57). The principle underlying non-linear least squares is exactly the same as ordinary least squares. If the sample is of size n and the data set consists of the three variables $w_i, p_{1i}, p_{2i}, i = 1, \ldots, n$, the problem is to minimise the following sum of squares with respect to the two parameters α and ρ:

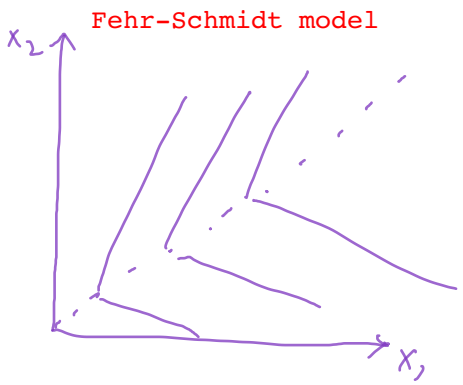
$$\sum_{i=1}^{n} \left[w_{1i} - \frac{p_{1i}^{\frac{\rho}{\rho-1}}}{p_{1i}^{\frac{\rho}{\rho-1}} + \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}} p_2^{\frac{\rho}{\rho-1}}} \right]^2 \tag{58}$$

The reason why non-linear least squares is required is that (58) is a non-linear function of the two parameters, and there is therefore no closed form expression for the solution to the minimisation problem, as there is when the model is linear. Instead, a numerical routine is used to locate the solution.

{tho} etc in {} are to be estimated

The STATA command which carries out non-linear least squares is `nl`. As with the regress command, it is possible to use the `vce(cluster i)` option to obtain cluster-robust standard errors. Another option we use here is "initial", in which starting values for the non-linear optimisation are provided. This option turns out to be essential - in the absence of starting values, estimation is not performed - although it is not necessary for the starting values to be particularly close to the solution.

The `nl` command with the two options just discussed, together with the results when



applied to the Andreoni & Miller (2002) data, are:

```
. nl (w1 = (p1^({rho}/({rho}-1)))/((p1^({rho}/({rho}-1)))  ///
> +(({aa}/(1-{aa}))^(1/({rho}-1)))*(p2^({rho}/({rho}-1)))), ///
> initial(rho 0.0 aa 0.5) vce(cluster i)
(obs = 1510)

Iteration 0:  residual SS =  122.2299
Iteration 1:  residual SS =  115.4766
Iteration 2:  residual SS =  115.4615
Iteration 3:  residual SS =  115.4615
Iteration 4:  residual SS =  115.4615

Nonlinear regression                                  Number of obs =      1510
                                                      R-squared     =    0.8804
                                                      Adj R-squared =    0.8798
                                                      Root MSE      =  .2767056
                                                      Res. dev.     =  403.0932

                                    (Std. Err. adjusted for 176 clusters in i)
------------------------------------------------------------------------------
             |               Robust
          w1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /rho |    .272248   .0479813     5.67   0.000     .1775515    .3669445
         /aa |   .6918387   .0150264    46.04   0.000     .6621824     .721495
------------------------------------------------------------------------------

.
. nlcom sigma: 1/(1- _b[rho:_cons])

       sigma:  1/(1- _b[rho:_cons])

------------------------------------------------------------------------------
          w1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sigma |   1.374095   .0905952    15.17   0.000     1.196531    1.551658
------------------------------------------------------------------------------
```

Following the `nl` command, we apply the `nlcom` command to deduce an estimate of the elasticity of substitution, σ, using the formula (56). We see that this estimate is 1.37 and the confidence interval provides evidence that the elasticity of substitution is larger than one. This indicates that the subjects in this sample, overall, attach somewhat more importance to efficiency than to equality in payoffs. The estimate of α is 0.692. This may be interpreted as the proportion of the allocation that the individual would take for themselves in a situation of equal prices. Being significantly greater than 0.5 (again based on the confidence interval), this estimate indicates that subjects are relatively selfish.

4.3 Estimation of Social Preference Parameters Using Discrete Choice Models

4.3.1 The setting

In this section, we once again pursue the goal of estimating the parameters of a utility function involving own payoff and other's payoff. However, the experimental approach is very different. Here, we are asking subjects to choose between hypothetical allocations. The approach is very similar to that of Engelmann & Strobel (2004).

The task faced by a subject is to choose between three different (hypothetical) allocations, A, B and C, say, of the type shown in Table 3. Importantly, the subject is

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

Table 3: An example of three hypothetical allocations. The survey respondent is given the identity of "Person 2".

always given the identity of "Person 2".

The three allocations in the example of Table 3 all have both attractions and drawbacks, and it is reasonable to expect people to divide between the three when asked to choose between them. A selfish individual would prefer allocation A, since, given their assumed identity as "Person 2", allocation A gives them a pay-off of 8 which is higher than the pay-off they would recieve under B or C (6 and 7 respectively).

However, as we have seen many times in this course, not all individuals are selfish, and some are highly altruistic. An individual who is "inequity-averse" is likely to choose allocation B, since equality is perfect under this allocation, with all three persons receiving the same pay-off.

In fact, there are different types of inequity aversion. It is useful to make a distinction between "disadvantageous" and "advantageous" inequality – a distinction highlighted by Fehr & Schmidt (1999). To see the difference, we continue with the example in Table 3, and consider a situation in which allocation B is not available, and individuals are required to choose between A and C. Both A and C are unequal allocations. An individual who is averse to "disadvantageous" inequality is likely to choose A over C, since C is an allocation in which they (Person 2) are *disadvantaged* relative to Person 1. An individual who is averse to "advantageous" inequality is likely to choose C over A, since A is an allocation in which they (Person 2) are *advantaged* relative to Person 3.

Some individuals are neither self-interested, nor overly concerned about inequity, but instead are motivated by *efficiency* considerations. Such individuals simply look for the allocation that is most efficient in terms of generating the highest total pay-off for the group. We see that when faced with the allocations in Table 3, such an individual would choose allocation C since this gives the highest total pay-off of 24.

Of course, it is unlikely that individuals are motivated by only one of the concerns described above. It is more likely that all of the concerns matter, perhaps to varying degrees. Hence we set out to estimate a utility function involving all of the concerns, and applying to all individuals. The parameter estimates will convey the importance of each concern.

4.3.2 Formalising the criteria for choosing between allocations

Let x_{jk} be the payoff to person k in allocation j. Each allocation has the following attributes.

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

Efficiency is simply the sum of the payoffs over all persons, regardless of the distribution. For the example in Table 3, the efficiency attributes for the three allocations are:

$$EFF_A = 20; EFF_B = 18; EFF_C = 24$$

2. Minimax: $MM_j = min(x_{jk}, k = 1, 2, 3)$

Minimax is the smallest payoff in the allocation. If an individual uses this as a criterion, they are displaying an extreme form of inequality aversion since they are only ever concerned with the welfare of the worst-off person. For the example in Table 3, the minimax attributes for the three allocations are:

$$MM_A = 4; MM_B = 6; MM_C = 7$$

3. Self: $SELF_j = x_{j2}$

"Self" is the decision-maker's own payoff. Remember that the decision-maker is assumed to take the role of Person 2; hence the definition of "self" as the payoff of Person 2 in the allocation. An individual using "self" as their criterion is clearly a selfish individual since they are concerned with their own welfare and have no regard for the welfare of any other individual. For the example in Table 3, the "self" attributes for the three allocations are:

$$SELF_A = 8; SELF_B = 6; SELF_C = 7$$

The final two attributes that we consider derive from the well known "Fehr-Schmidt utility function" (Fehr & Schmidt 1999). If there are n persons in total, this utility function, for person i, is:

$$u_i = x_i - \alpha_i \frac{\sum_{k \neq i} \max(x_k - x_i, 0)}{n - 1} - \beta_i \frac{\sum_{k \neq i} \max(x_i - x_k, 0)}{n - 1} \tag{59}$$

The interpretation of (59) is as follows. Individual i's utility is given by their own payoff, penalised by the presence of two different types of inequality. The second term on the right-hand side is the term that adjusts for "disadvantageous inequality"; that is, inequality resulting from others receiving *higher* payoffs than themselves. The final term is the term that adjusts for "advantageous inequality"; that is, inequality resulting from others receiving *lower* payoffs than themselves. It is usually assumed that both types of inequality are undesirable, so the two parameters α_i and β_i, respectively referred to as individual i's coefficients of disadvantageous and advantageous inequity aversion, are both expected to be positive. It is also assumed that individuals care more about disadvantageous inequity than about advantageous inequity, so that $\alpha_i > \beta_i$. The denominators of the two inequity terms are present simply to ensure that the measures to not rise with the number of persons in the "economy".

In the case $n = 3$, the two measures of inequality just described become:

4a. (absence of) Disadvantageous inequality: $FSD_j = -\frac{1}{2}\sum_{k\neq 2} max(x_{jk} - x_{j2}, 0)$
4b. (absence of) Advantageous inequalty: $FSA_j = -\frac{1}{2}\sum_{k\neq 2} max(x_{j2} - x_{jk}, 0)$

FSD and FSA are the attributes in which we are interested. In these acronyms, "FS" stands for Fehr-Schmidt, and "D" and "A" for disadvantageous and advantageous respectively. Defining them with a negative sign allows them to be treated as positive attributes, that is, quantities that an inequity-averse individual would seek to maximise.

For the example in Table 3, the attributes FSD and FSA for the three allocations are:

$$FSD_A = 0; \quad FSD_B = 0; \quad FSD_C = -\frac{3}{2}$$
$$FSA_A = -2; \quad FSA_B = 0; \quad FSA_C = 0$$

4.3.3 Data

Simulated data closely resembling Engelmann and Strobel's data is contained in the file **ES_sim**. Note that there are three rows for each subject, one row for each allocation (this is known as a "long" data set). The attributes are named as in Section 4.3.2. y is a binary variable indicating which of the three allocations is chosen (1 if chosen; 0 if not chosen).

4.3.4 The conditional logit model (CLM)

Let us henceforth use i to index an individual in the data set. Each individual chooses one from $J = 3$ possible allocations. Suppose that individual i's utility from choosing allocation j is given by:

$$U_{ij} = \alpha_1 FSD_{ij} + \alpha_2 FSA_{ij} + \alpha_3 EFF_{ij} + \alpha_4 MM_{ij} + \epsilon_{ij} = z'_{ij}\alpha + \epsilon_{ij} \qquad (60)$$

Note that the attribute variables have both i and j subscripts, indicating that different individuals face allocations with different sets of attributes. Note also that there is no intercept in (60). This is because an intercept parameter would not be identified, because, as is well known, adding a constant to a utility function does not alter implied behaviour. For convenience, we have collected the list of attributes together in the vector z_{ij} and the associated parameters into the vector α. The term $z'_{ij}\alpha$ may be referred to as the deterministic component of utility, and ϵ_{ij} as the random component.

We then assume that each individual chooses the allocation that yields the highest utility. Formally, the *observed* decision variable is y_{ij}, and:

$$\begin{aligned} y_{ij} &= 1 \quad \text{if} \quad U_{ij} = max\,(U_{i1}, U_{i2} \ldots U_{iJ}) \\ y_{ij} &= 0 \quad \text{otherwise} \end{aligned} \tag{61}$$

We then need to consider what is the probability that individual i will choose allocation j. This depends on what is assumed about the distribution of the random component of utility.

$$\begin{aligned} y_{ij} = 1 &\Leftrightarrow z'_{ij}\alpha + \epsilon_{ij} > z'_{ik}\alpha + \epsilon_{ik} \quad \forall k \neq j \\ &\Leftrightarrow \epsilon_{ik} - \epsilon_{ij} < z'_{ij}\alpha - z'_{ik}\alpha \quad \forall k \neq j \end{aligned} \tag{62}$$

For convenience, it is assumed that the ϵ_{ij}s are independent and identically distributed (i.i.d.) with type I extreme value distribution (also known as the Gumbel distribution), defined by the density function:

$$f(\epsilon) = \exp\left(-\epsilon - \exp(-\epsilon)\right) \quad -\infty < \epsilon < \infty \tag{63}$$

and the distribution function:

$$F(\epsilon) = \exp\left(-\exp(-\epsilon)\right) \quad -\infty < \epsilon < \infty \tag{64}$$

It can be shown (Maddala 1983) that if the ϵ_{ij}s are i.i.d. and have the distribution defined in (63) and (64), then the probability of the event defined in (62), that is, the probability of allocation j being chosen by individual i, is given by:

$$P(y_{ij} = 1) = \frac{exp(z'_{ij}\alpha)}{\sum_{k=1}^{J} exp(z'_{ik}\alpha)} \tag{65}$$

The model defined by (65) is known as the *conditional logit model*. The likelihood contribution associated with individual i is given by:

$$L_i(\alpha) = \frac{\sum_{k=1}^{J} y_{ik} exp(z'_{ik}\alpha)}{\sum_{k=1}^{J} exp(z'_{ik}\alpha)} \tag{66}$$

from which the sample log-likelihood is obtained as:

$$LogL(\alpha) = \sum_{i=1}^{n} \ln L_i(\alpha) \tag{67}$$

4.3.5 Results

The parameters contained in the vector α in (60) are estimated by maximisation of the likelihood function defined in (66) and (67) using the following STATA command:

```
. asclogit y FSD FSA EFF MM, case( i) alternatives(j) noconstant
```

The `as` at the beginning of the command name stands for "alternative-specific".

The output looks like this:

```
Iteration 0:   log likelihood = -317.10088
Iteration 1:   log likelihood = -308.55197
Iteration 2:   log likelihood = -308.51212
Iteration 3:   log likelihood = -308.51212

Alternative-specific conditional logit          Number of obs      =        990
Case variable: i                                Number of cases    =        330

Alternative variable: t                         Alts per case: min =          3
                                                               avg =        3.0
                                                               max =          3

                                                  Wald chi2(4)     =      80.96
Log likelihood = -308.51212                       Prob > chi2      =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
t            |
         FSD |   .3267221   .1405881     2.32   0.020     .0511745    .6022697
         FSA |   .3447768   .1688655     2.04   0.041     .0138065    .6757472
         EFF |   .1879009   .0714842     2.63   0.009     .0477943    .3280074
          MM |   .0804075   .0895162     0.90   0.369    -.0950409     .255856
------------------------------------------------------------------------------
```

The two inequity aversion attributes have been included, as well as the efficiency and minimax attributes. It appears (from this simulated data set) that subjects display both types of inequity aversion: both FSD and FSA have significantly positive effects on utility. Efficiency appears to be even more important: the coefficient of EFF is strongly significantly positive. The minimax attribute (MM) does not appear to be important.

Recall that another attribute was SELF defined as payoff to self. It turns out that the experimental design is such that when SELF is added to the above model, the problem of perfect multicollinearity arises resulting in a failure to estimate the effect of this variable. This is why SELF has been excluded.

4.3.6 The effect of subject characteristics

It is reasonable to expect individuals to value different criteria differently. In previous parts, we have allowed for this type of difference between subjects by building *unobserved heterogeneity* into models. Here we will demonstrate an alternative approach: *observed heterogeneity*. Observed heterogeneity refers to the situation in which differences between subjects can be explained by differences in subject characteristics, perhaps the most obvious being gender.

The way in which subject characteristics are introduced in the CLM is through *interactions* with the attribute variables. Let the variable $male_i$ be a dummy variable taking the value 1 if subject i is male. An important point is that subject characteristics have only an i subscript, as distinct from the attributes which have both an i and a j subscript.

Let us extend the utility function (60) with two additional terms:

$$U_{ij} = \alpha_1 FSD_{ij} + \alpha_2 FSD_{ij} \star \text{male}_i + \alpha_3 FSA_{ij} + \alpha_4 FSA_{ij} \star \text{male}_i + \alpha_5 EFF_{ij} + \alpha_6 MM_{ij} + \epsilon_{ij} \quad (68)$$

The two additional terms are interactions between the dummy variable male and the two inequity attributes. When these two variables are included in the model, the results are as follows:

```
Alternative-specific conditional logit          Number of obs      =        990
Case variable: i                                Number of cases    =        330

Alternative variable: j                         Alts per case: min =          3
                                                               avg =        3.0
                                                               max =          3

                                                  Wald chi2(6)     =      85.42
Log likelihood =  -299.6794                       Prob > chi2      =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
j            |
         FSD |   .1907648   .1552983     1.23   0.219    -.1136143     .495144
    male_FSD |   .2535549   .1281861     1.98   0.048     .0023147     .504795
         FSA |   .5649655   .1879811     3.01   0.003     .1965293    .9334017
    male_FSA |  -.5760542    .192775    -2.99   0.003    -.9538863   -.1982221
         EFF |   .1606768   .0741216     2.17   0.030     .0154012    .3059525
          MM |   .1170375    .091562     1.28   0.201    -.0624207    .2964958
------------------------------------------------------------------------------
```

Here (not forgetting that the data is simulated), we see a very interesting result: the significantly positive coefficient on the interaction male_FSD indicates that males exhibit more aversion to disadvantageous inequity than do females; the significantly negative coefficient on male_FSA indicates that females exhibit more aversion to advantageous inequity than do males.

5 Dealing with heterogeneity: Finite Mixture Models

5.1 Introduction

Finite mixture models, or just mixture models, are a class of model that offer a means of separating subjects into different types. Different types do not only exhibit different behaviour; *the processes giving rise to behaviour* also differ between types. These models are labelled as "finite" mixture models because a finite number of types is being assumed. An "infinite" mixture model, if such a label were used, would correspond to a random coefficient model, or random effects model, in which it is assumed that there is continuous variable in some parameter indexing behavioural type.

Finite mixture models are very important in experimetrics. This is because it is becoming ever more widely accepted that different subjects are motivated in different ways, and to assume that all subjects operate according to one model is to disrespect these differences. Often average behaviour is tracked and interpreted in terms of the behaviour of a typical subject. However, if there are different types of subject operating according to different decision processes, it is quite possible that average behaviour will not be a close representation of the actual behaviour of any of the subjects under study.

There is more than one possible approach to the estimation of finite mixture models. The approach adopted here is as follows. Firstly, on the basis of economic theory, the total number of types in the population is decided, and a label is assigned to each. Then a parametric model is specified for the behaviour of each type. The parameters of these various models are estimated altogether, along with the "mixing proportions" - parameters revealing the proportion of the population who is of each type. Once the model has been estimated, we can return to the data in order to determine the posterior probability of each individual subject being of each type. Note that there is no claim to be able to identify any individual subject as belonging to any particular type with certainty, although, in situations where data is informative, posterior type-probabilities can be very close to one.

We commence with a simple, somewhat contrived, example: a mixture of two normal distributions. Then we shall progress to a more realistic example of guesses in the "Beauty Contest game". Finally, we shall consider the more complex example of giving in a public goods experiment.

5.2 Mixture of two normal distributions

Consider the variable y contained in the file **mixture_sim**. There are 1000 observations. A histogram of the variable y is shown in Figure 16. The distribution appears to be a mixture of two bell-shaped (i.e. probably normal) distributions, one with a mean around 3, the other with a mean around 6.

If y represents the decision made by subjects in an experiment (on this occasion, let us not concern ourselves about what decision is actually being made), we might

```
hist y
```

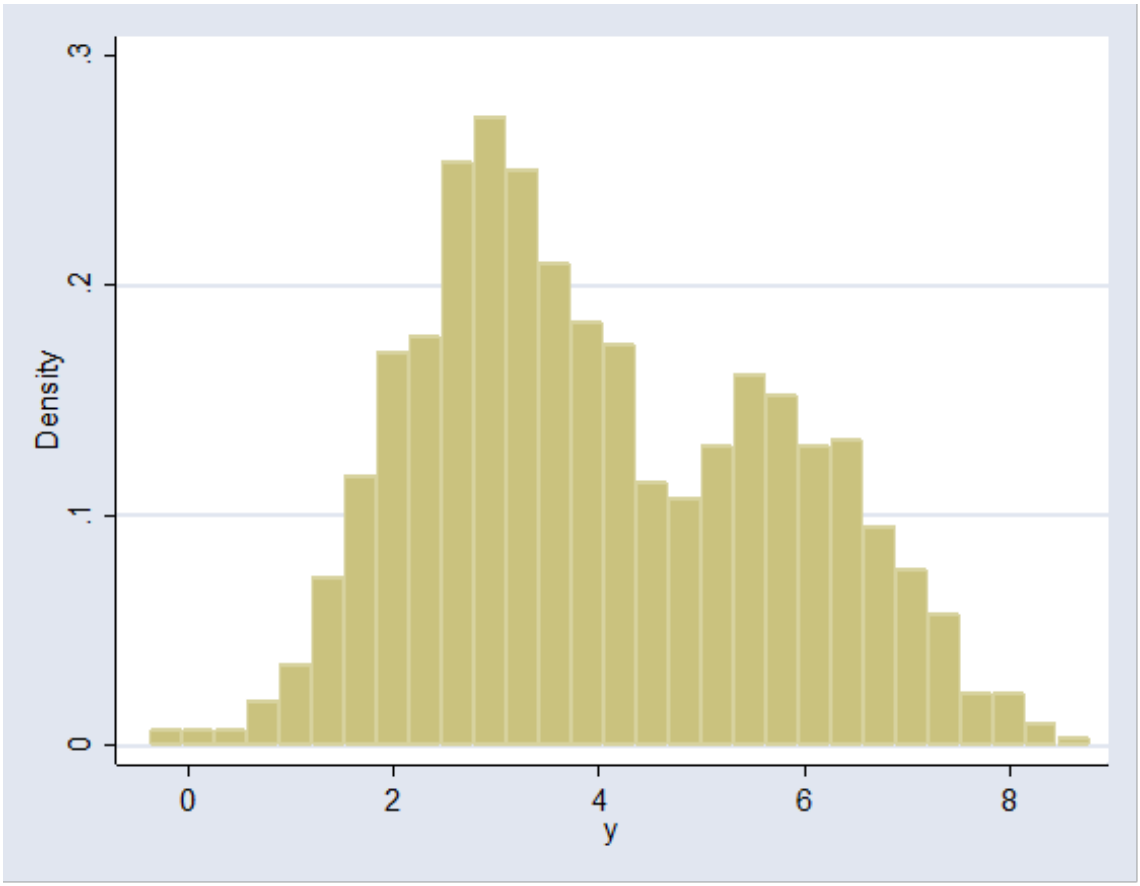


Figure 16: A histogram of the variable y in the data set **mixture_sim**.

say that subjects are of two types, and we would set about estimating the following mixture model:

$$\begin{aligned} &\text{type 1:} && N\left(\mu_1, \sigma_1^2\right) \\ &\text{type 2:} && N\left(\mu_2, \sigma_2^2\right) \\ &\text{mixing proportions:} && p(\text{type1}) = p \quad p(\text{type2}) = 1 - p \end{aligned}$$

The mixing proportions represent the proportions of the population who are of each type. Note that there are five parameters to be estimated: $\mu_1, \sigma_1, \mu_2, \sigma_2$ and p.

The density associated with a particular value of y, conditional on the subject being of type 1, is:

$$f\left(y|\text{type1}\right) = \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) \tag{69}$$

and the density conditional on being type 2 is:

$$f\left(y|\text{type2}\right) = \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right) \tag{70}$$

Therefore the *marginal* density associated with an observation is obtained by com-

bining (69) and (70) with the mixing proportions:

$$f(y;\mu_1,\sigma_1,\mu_2,\sigma_2,p) = p \times \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + (1-p) \times \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right) \tag{71}$$

Equation (71) is used as the likelihood contribution for each observation.

The sample log-likelihood, based on a sample $y_1, y_2, \ldots, y_n$ is given by:

$$LogL = \sum_{i=1}^{n} \ln f(y_i;\mu_1,\sigma_1,\mu_2,\sigma_2,p) \tag{72}$$

Maximising (72) with respect to μ_1, σ_1, μ_2, σ_2, and p, we will obtain MLEs of these 5 parameters.

5.2.1 Posterior type probabilities

Having estimated a mixture model, one obvious thing to do is to compute the posterior probabilities of each subject being of each type. This involves Bayes' rule. For example, the posterior type 1 probability, given an observation y is:

$$\begin{aligned} P(\text{type1}|y) &= \frac{f(y|\text{type1})P(\text{type1})}{f(y|\text{type1})P(\text{type1}) + f(y|\text{type2})P(\text{type2})} \\ &= \frac{p \times \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right)}{p \times \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + (1-p) \times \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)} \end{aligned} \tag{73}$$

5.2.2 The estimation program

A STATA program which estimates the model and then computes and plots posterior probabilities is shown below. Table 4 provides a correspondence between components of the likelihood function (72) and the names used in the program.

Component of $LogL$	STATA name
μ_1,μ_2	mu1, mu2
σ_1, σ_2	sig1, sig2
p	p
$f(y\|\text{type1}) = \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right)$	f1
$f(y\|\text{type2}) = \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)$	f2
$\ln[f(y)] = \ln\left[p \times \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + (1-p) \times \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)\right]$	logl
$P(\text{type1}\|y)$	postp1
$P(\text{type 2}\|y)$	postp2

Table 4: Components of $LogL$ and corresponding STATA names.

The annotated code is as follows. One important point is that the variable y is a "global" variable, because it exists both inside and outside the likelihood-evaluation program. This is why quotation marks are not used when y is used within the program.

```
program drop _all

*  LIKELIHOOD EVALUATION PROGRAM STARTS HERE:

program define mixture
args logl mu1 sig1 mu2 sig2 p
tempvar f1 f2

* GENERATE TYPE-CONDITIONAL DENSITIES:

quietly gen double `f1'=(1/`sig1')*normalden((y-`mu1')/`sig1')
quietly gen double `f2'=(1/`sig2')*normalden((y-`mu2')/`sig2')

* COMBINE TYPE-CONDITIONAL DENSITIES WITH MIXING PROPORTIONS
* TO GENERATE MARGINAL DENSITY.  THIS IS THE FUNCTION THAT
* NEEDS TO BE MAXIMISED WHEN SUMMED OVER THE SAMPLE:

quietly replace `logl'=ln(`p'*`f1'+(1-`p')*`f2')

* GENERATE THE POSTERIOR TYPE PROBABILITIES, AND MAKE THEM
* AVAILABLE OUTSIDE THE PROGRAM:

quietly replace postp1=`p'*`f1'/(`p'*`f1'+(1-`p')*`f2')
quietly replace postp2=(1-`p')*`f2'/(`p'*`f1'+(1-`p')*`f2')

quietly putmata postp1, replace
quietly putmata postp2, replace

end

* END OF LIKELIHOOD EVALUATION PROGRAM

* READ DATA:

use mixture_sim, clear

* INITIALISE TWO POSTERIOR PROBABILITY VARIABLES:

gen postp1=.
gen postp2=.

* SPECIFY STARTING VALUES, AND APPLY ML:

mat start=(3, 1.5, 6, 1.5, .5)
ml model lf mixture /mu1 /sig1 /mu2 /sig2 /p
ml init start, copy
ml maximize

* EXTRACT POSTERIOR TYPE PROBABILITY, AND PLOT THEM AGAINST y:

drop postp1 postp2
getmata postp1
getmata postp2

sort y
line  postp1 postp2 y , lpattern(l -)
```

As usual, it is necessary to specify starting values for the parameters being estimated. These are stored in the vector "start". In this case, the starting values have been obtained by examination of the histogram of y (see fig 16). In other situations, starting values are obtained using simple estimation methods such as linear regression.

5.2.3 Results

The results from executing the code presented in sect 5.2.2 above are as follows:

```
                                                Number of obs   =       1000
                                                Wald chi2(0)    =          .
Log likelihood = -1908.2805                     Prob > chi2     =          .

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
       _cons |   2.981757   .0743116    40.13   0.000     2.836109    3.127405
-------------+----------------------------------------------------------------
eq2          |
       _cons |   1.014725   .0499721    20.31   0.000     .9167818    1.112669
-------------+----------------------------------------------------------------
eq3          |
       _cons |   5.950353   .1158028    51.38   0.000     5.723384    6.177322
-------------+----------------------------------------------------------------
eq4          |
       _cons |   .9768525   .0721166    13.55   0.000     .8355064    1.118198
-------------+----------------------------------------------------------------
eq5          |
       _cons |   .6494311   .0296983    21.87   0.000     .5912235    .7076387
------------------------------------------------------------------------------
```

We see that the estimates of the five parameters (with standard errors) are:

$$\begin{aligned}\hat{\mu}_1 &= 2.982(0.074)\\ \hat{\sigma}_1 &= 1.015(0.050)\\ \hat{\mu}_2 &= 5.950(0.116)\\ \hat{\sigma}_2 &= 0.977(0.072)\\ \hat{p} &= 0.649(0.030)\end{aligned}$$

Hence we see that 64.9% of the population comes from the distribution $N(2.982, 1.015^2)$, while the remaining 35.1% comes from $N(5.950, 0.977^2)$.

However, when considering any particular observation, we might not be certain which one of the two distributions it comes from. This is why the posterior probabilities are useful. Note that the variables containing the posterior probabilities (postp1 and postp2) are generated inside the likelihood evaluation program. In order to extract these variables, mata commands are required. The `putmata` command is used from within the program, and the getmata command is used outside it.

Below, we show the plot of the posterior probabilities against y.

```
sort y
line  postp1 postp2 y , lpattern(l -)
```

This graph tells us that: observations below 3 are almost certain to be from the first distribution; observations greater than 6 are almost certain to come from the second distribution. For observations between 3 and 6, we cannot know with confidence which distribution applies. For an observation with $y = 4.70$, both distributions are equally likely to apply.

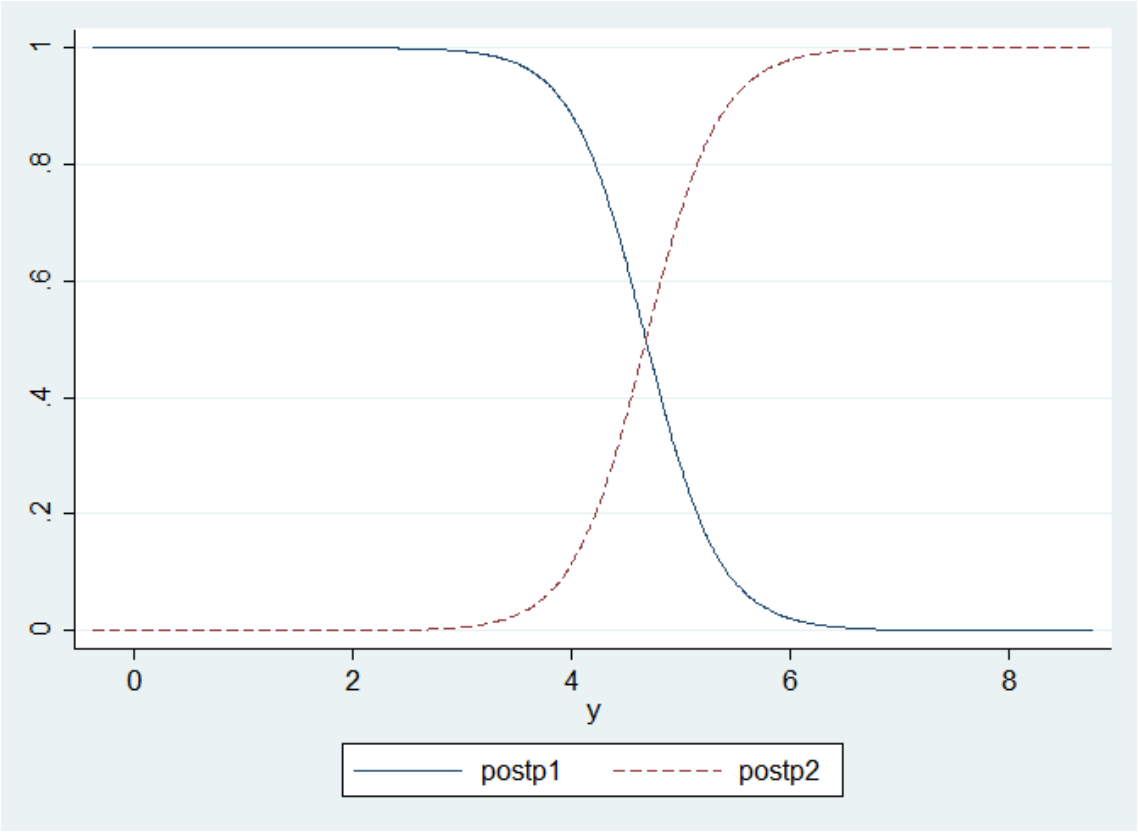


Figure 17: Posterior probabilities of type 1 and type 2 in the mixture data.

5.3 The fmm command in STATA

STATA 15 command

fmm (for "finite mixture model") is a STATA command that directly estimates mixture models of the type considered in Sect. 5.2.

The required syntax is:

```
fmm 2: regress y
```

The number before the colon lets STATA know that there are 2 types in the mixture. After the colon, the model to be estimated for each type is specified. Note that this is a regression model with an intercept only (which is equivalent to estimating a mean). The results from this command are as follows. Note that these results are equivalent to the results obtained using ml in Sect. 5.2, except that some parameters are named differently.

```
. fmm 2: regress y

Fitting class model:

Iteration 0:   (class) log likelihood = -693.14718
Iteration 1:   (class) log likelihood = -693.14718

Fitting outcome model:

Iteration 0:   (outcome) log likelihood = -1340.7846
Iteration 1:   (outcome) log likelihood = -1340.7846

Refining starting values:

Iteration 0:   (EM) log likelihood =  -2114.989
```

```
Iteration 1:   (EM) log likelihood = -2144.1684
Iteration 2:   (EM) log likelihood =  -2155.951
Iteration 3:   (EM) log likelihood = -2159.9264
Iteration 4:   (EM) log likelihood = -2159.9464
Iteration 5:   (EM) log likelihood = -2157.8613
Iteration 6:   (EM) log likelihood = -2154.6472
Iteration 7:   (EM) log likelihood = -2150.8481
Iteration 8:   (EM) log likelihood = -2146.7758
Iteration 9:   (EM) log likelihood = -2142.6116
Iteration 10:  (EM) log likelihood = -2138.4622
Iteration 11:  (EM) log likelihood = -2134.3904
Iteration 12:  (EM) log likelihood = -2130.4335
Iteration 13:  (EM) log likelihood = -2126.6137
Iteration 14:  (EM) log likelihood = -2122.9441
Iteration 15:  (EM) log likelihood =  -2119.432
Iteration 16:  (EM) log likelihood = -2116.0816
Iteration 17:  (EM) log likelihood = -2112.8942
Iteration 18:  (EM) log likelihood = -2109.8699
Iteration 19:  (EM) log likelihood = -2107.0071
Iteration 20:  (EM) log likelihood = -2104.3034
Note: EM algorithm reached maximum iterations.

Fitting full model:

Iteration 0:   log likelihood = -1909.8137
Iteration 1:   log likelihood = -1908.4031
Iteration 2:   log likelihood = -1908.2811
Iteration 3:   log likelihood = -1908.2805
Iteration 4:   log likelihood = -1908.2805

Finite mixture model                            Number of obs     =      1,000
Log likelihood = -1908.2805

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Class      |  (base outcome)
-------------+----------------------------------------------------------------
2.Class      |
       _cons |  -.6165402    .130444    -4.73   0.000    -.8722058   -.3608746
------------------------------------------------------------------------------

Class          : 1
Response       : y
Model          : regress

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
       _cons |   2.981758   .0743115    40.13   0.000      2.83611    3.127406
-------------+----------------------------------------------------------------
     var(e.y)|   1.029668   .1014158                      .848905    1.248921
------------------------------------------------------------------------------

Class          : 2
Response       : y
Model          : regress

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
       _cons |   5.950353   .1158024    51.38   0.000     5.723385    6.177322
-------------+----------------------------------------------------------------
     var(e.y)|   .9542398   .1408942                      .7144585    1.274495
------------------------------------------------------------------------------
```

Following the fmm command, The postestimation command predict may be used to obtain the posterior type probabilities, as follows:

```
predict post1 , pos eq(component1)
predict post2 , pos eq(component2)
```

The posterior probabilities (post1 and post2) thus obtained are identical to those obtained in Sect. 5.2.3.

5.4 Application 1: A Level-k Model for the Beauty Contest Game

Nagel (1995)'s "p-Beauty Contest game" takes the following form. Each player chooses a whole number between 0 and 100. In the case in which $p = 2/3$, the winner is the player whose number is closest to 2/3 of the average for the entire group.

Let us imagine that this game has been played in a large lecture theatre, with 500 players. Simulated data is contained in the file **beauty_sim**. The distribution of simulated guesses is shown in Fig. 18. We see that the distribution is multi-modal, with one clear modes at around 33 and and another around 22. Note that there is another mode close to zero.

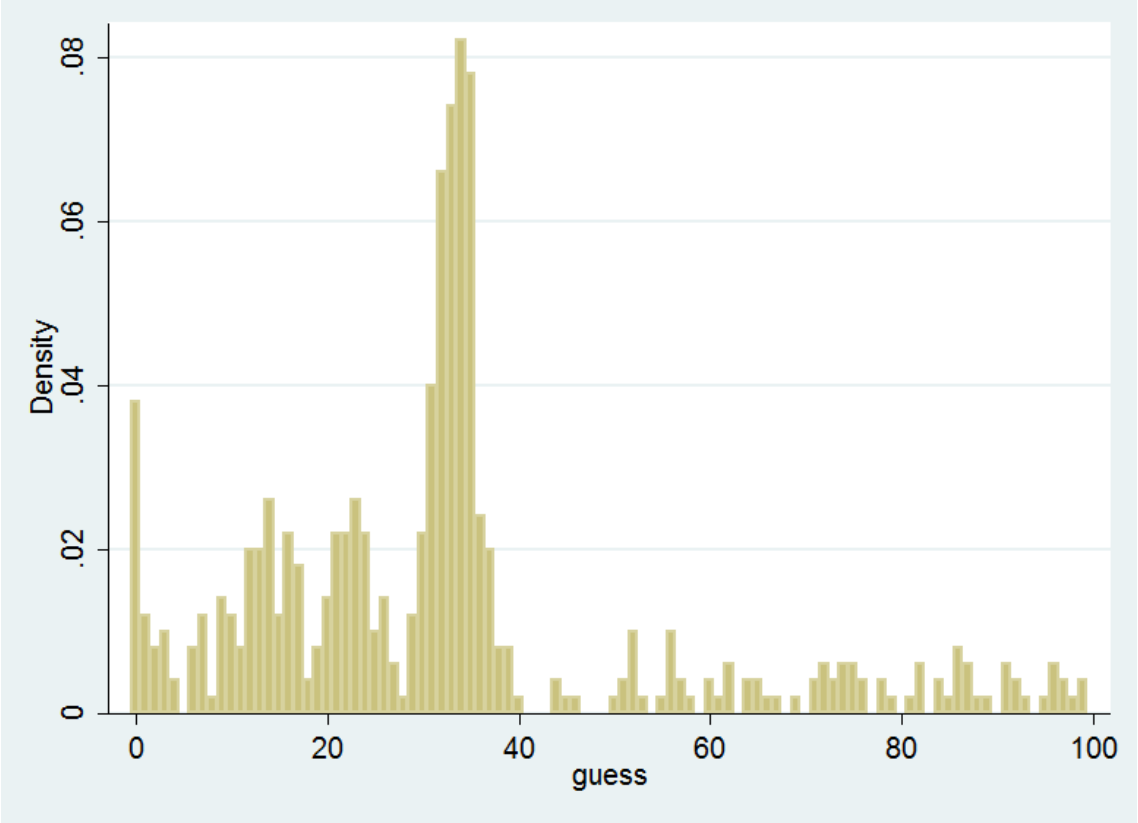


Figure 18: The distribution of simulated guesses for the beauty contest game.

A popular approach to modelling behaviour in this game is the level-k model, a simple version of which is as follows. We first assume that there are a group of individuals in the population who select a number completely at random, from a uniform distribution on $[0, 100]$. Strictly speaking, we further assume that this draw is rounded to the nearest integer. These individuals are labelled "level-0 reasoners". Then, there is a group who assume that all other players are level-0 reasoners, inferring that the mean guess will be around 50, and therefore that the best guess

is 33 (being the closest integer to 2/3 of 50). These players are labelled "level-1 reasoners". Next, there is a group who assume that all others are level-1, with a mean guess of 33, so that the best guess for this type of individual is 22; these are level-2 reasoners. This sequence continues. Level-3 reasoners will guess 15. Level-4 reasoners will guess 10, and so on.

Note that if every player had immaculate powers of reasoning, they would all supply a guess of 0, and they would all be correct and share the prize. However, needless to say, this is not what happens when the game is played with real subjects.

The estimation problem is to use the data shown in the histogram to estimate the proportion of the population who are at each level of reasoning. We require a parametric model. We assume that there are a finite number $(J+1)$ of types, and the maximum level of reasoning is level J.

In practice, some subjects do go straight to the Nash Equilibrium of zero, so it is sensible to allow for a "naïve-Nash-type" whose best guess is assumed to be zero. They are labelled as "naïve" simply because, although their behaviour corresponds to the Nash prediction, they are very unlikely to win the game. We will assign the level-J label to this type.

Apart from Level-0 reasoners, who are assumed to choose from a uniform distribution, we assume that an individual's choice is the best guess for an individual of their type, plus a random normally distributed error with mean zero and standard deviation σ. That is, we assume that if y_j^* is the best guess for type j, then the actual guess (y) will be determined by:

$$(y|\text{type } j) = y_j^* + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \quad j = 1, ..., J \ . \tag{74}$$

These assumptions give us the conditional density functions for each type:

$$\begin{aligned} f(y|L_0) &= 1/100 && 0 \le y \le 100 \\ f(y|L_j) &= \frac{1}{\sigma}\phi\left(\frac{y - y_j^*}{\sigma}\right) && 0 \le y \le 100 \quad j = 1, ..., J \ . \end{aligned} \tag{75}$$

We also assume that the population is made up of the $J+1$ types with mixing proportions $p_0, p_1, \cdots, p_J$. Combining the mixing proportions with the conditional densities (75) gives us the sample log-likelihood (for a sample of guesses y_i, $i = 1, ..., n$):

$$\text{LogL} = \sum_{i=1}^{n} \ln\left[\frac{p_0}{100} + \sum_{j=1}^{J} p_j \frac{1}{\sigma}\phi\left(\frac{y_i - y_j^*}{\sigma}\right)\right] \ . \tag{76}$$

We set $J = 5$, and the "best guesses" are $y_1^* = 33$, $y_2^* = 22$, $y_3^* = 15$, $y_4^* = 10$, $y_5^* = 0$. Note that the best guess for level 5 is zero, because, as noted above, we assign level J to "naïve-Nash" subjects. The code required to maximise the log-likelihood function is:

```
program define beauty_mixture
args lnf  p1 p2 p3 p4 p5 sig
tempvar  f0 f1 f2 f3 f4 f5 l
```

```
quietly{

gen double `f0'=0.01
gen double `f1'=(1/`sig')*normalden((y-33.5)/`sig')
gen double `f2'=(1/`sig')*normalden((y-22.4)/`sig')
gen double `f3'=(1/`sig')*normalden((y-15.0)/`sig')
gen double `f4'=(1/`sig')*normalden((y-10.1)/`sig')
gen double `f5'=(1/`sig')*normalden((y-0)/`sig')

gen double `l'=(1-`p1'-`p2'-`p3'-`p4'-`p5')*`f0'   ///
+`p1'*`f1'+`p2'*`f2'+`p3'*`f3'+`p4'*`f4'+`p5'*`f5'

replace postp0=(1-`p1'-`p2'-`p3'-`p4'-`p5')*`f0'/`l'
replace postp1=`p1'*`f1'/`l'
replace postp2=`p2'*`f2'/`l'
replace postp3=`p3'*`f3'/`l'
replace postp4=`p4'*`f4'/`l'
replace postp5=`p5'*`f5'/`l'

 replace `lnf'=ln((1-`p1'-`p2'-`p3'-`p4'-`p5')*`f0'+`p1'*`f1' ///
+`p2'*`f2'+`p3'*`f3'+`p4'*`f4'+`p5'*`f5')


 putmata postp0, replace
 putmata postp1, replace
 putmata postp2, replace
 putmata postp3, replace
 putmata postp4, replace
 putmata postp5, replace

}

end



gen postp0=.
gen postp1=.
gen postp2=.
gen postp3=.
gen postp4=.
gen postp5=.


mat start=(0.3, 0.4, 0.1, 0.1,0.05, 2)
ml model lf beauty_mixture /p1 /p2 /p3 /p4 /p5 /sig
ml init start, copy

ml maximize

nlcom p0: 1-_b[p1:_cons]-_b[p2:_cons]-_b[p3:_cons]-_b[p4:_cons]-_b[p5:_cons]

drop postp*

getmata postp0
getmata postp1
getmata postp2
getmata postp3
getmata postp4
getmata postp5

sort y

line  postp0 postp1 postp2 postp3 postp4 postp5 y , lpattern(- l l l l l)
```

Note that the five mixing proportions $p_1, ..., p_5$ are estimated, and then an estimate of p_0 is deduced using the delta method. The results are:

```
. mat start=(0.3, 0.4, 0.1, 0.1,0.05, 2)
. ml model lf beauty_mixture /p1 /p2 /p3 /p4 /p5 /sig
. ml init start, copy
. ml maximize

                                                  Number of obs   =        500
```

```
                                                Wald chi2(0)    =          .
Log likelihood = -1985.0613                     Prob > chi2     =          .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
p1           |
       _cons |   .3982665    .023804    16.73   0.000     .3516116    .4449213
-------------+----------------------------------------------------------------
p2           |
       _cons |   .1128533   .0163975     6.88   0.000     .0807148    .1449919
-------------+----------------------------------------------------------------
p3           |
       _cons |   .0898775   .0159347     5.64   0.000     .0586461     .121109
-------------+----------------------------------------------------------------
p4           |
       _cons |   .0462681   .0135852     3.41   0.001     .0196415    .0728946
-------------+----------------------------------------------------------------
p5           |
       _cons |   .0500939   .0117892     4.25   0.000     .0269876    .0732002
-------------+----------------------------------------------------------------
sig          |
       _cons |   1.929627   .1027345    18.78   0.000     1.728271    2.130982
------------------------------------------------------------------------------

.
. nlcom p0: 1-_b[p1:_cons]-_b[p2:_cons]-_b[p3:_cons]-_b[p4:_cons]-_b[p5:_cons]

          p0:  1-_b[p1:_cons]-_b[p2:_cons]-_b[p3:_cons]-_b[p4:_cons]-_b[p5:_cons
> ]

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p0 |   .3026407    .029052    10.42   0.000     .2456999    .3595815
------------------------------------------------------------------------------
```

We see that the mixing proportions and the standard deviation of the error term are estimated to be:

$$\begin{aligned}\hat{p}_0 &= 0.303(0.029)\\ \hat{p}_1 &= 0.398(0.024)\\ \hat{p}_2 &= 0.113(0.016)\\ \hat{p}_3 &= 0.090(0.016)\\ \hat{p}_4 &= 0.046(0.014)\\ \hat{p}_5 &= 0.050(0.012)\\ \hat{\sigma} &= 1.930(0.103)\end{aligned}$$

It appears that in this (simulated) data set, around 30% of subjects are estimated to be level-0. Of the remainder, in agreement with similar studies that use real data, the majority are divided between levels 1, 2 and 3. The proportion of "naïve Nash-types" is 0.05.

We may then consider the posterior type probabilities which have been generated in the usual way using Bayes' rule. It is sensible to plot these against the subject's guess. This is done in Fig. 19. The dashed curve represents the posterior probability of level-0. Note that this is close to 1 for any subject whose guess is greater than about 40. The other posterior probabilities peak in different positions, as expected. The curve peaking at 33 is the level-1 posterior probability; the one peaking at 22 is that for level-2; the one at 15 is for level-3; the one at 10 is for level 4. The curve

peaking at zero is for level-5, the "naïve Nash-type". The position of this last curve indicates that subjects whose guess is zero or a *very small positive number* may be categorised as "naïve Nash".

It is perhaps interesting that the dashed line (probability of being level-0) appears to reach peaks in the areas between the various "best guesses". If for example, if a subject's guess is 27 or 28 (i.e. roughly half way between the best guess for levels 1 and 2) they are neither likely to be level-1 or level-2, and are instead categorised as level-0, as indicated by the peak in the dashed curve in the vicinity of this point.

Finally, it is interesting to consider what the winning guess is. The mean of the simulated sample is 32.5, implying that the winning guess is 22. The vertical line drawn in fig 19 represents this winning guess. This happens to be the best guess for a level-2 type. Accordingly, we see that the probability of the winner being level-2 is around 0.86. We also see that the winner has a perhaps surprisingly high probability of around 0.14 of being level-0.

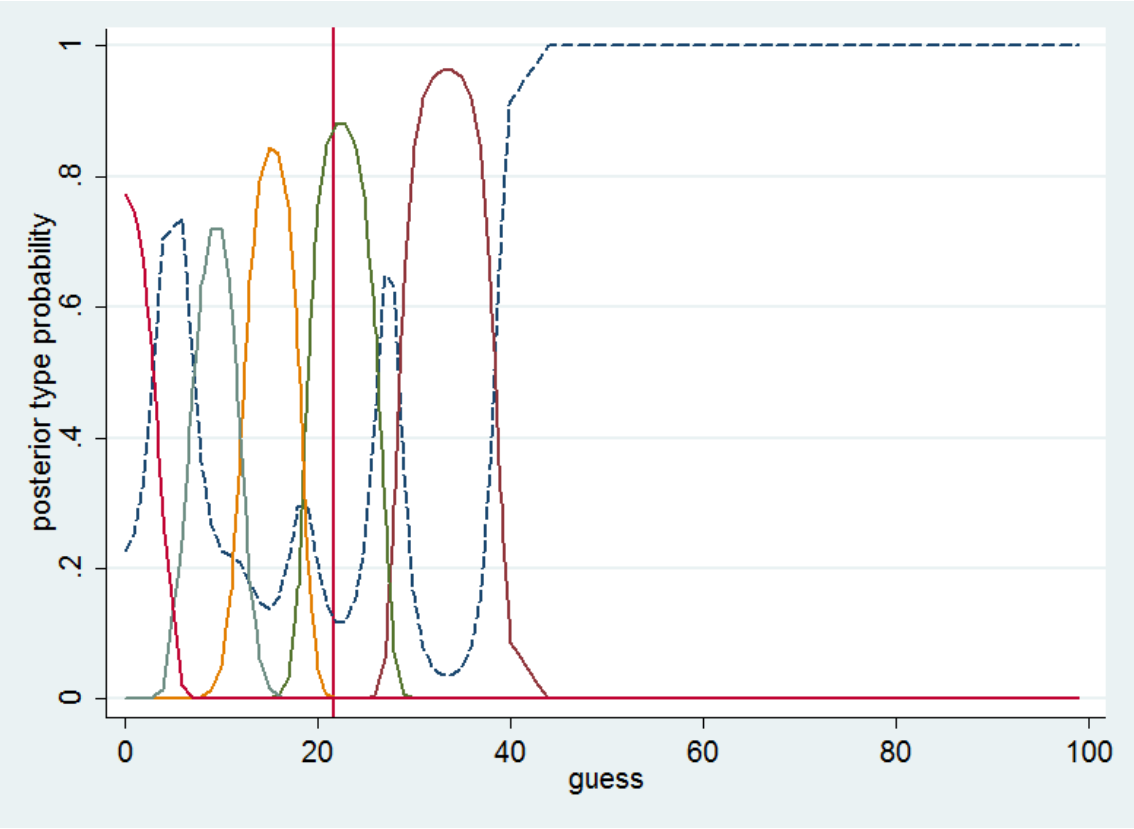


Figure 19: Posterior type probabilities in the level-k model.

5.5 Application 2: A public goods experiment

5.5.1 Background

In a typical public goods experiment, each subject has to divide an endowment between a public account and a private account. Total contributions to the public account are then multiplied up by some factor and divided equally between the

group of participants.

If everyone is a selfish agent, the game has a unique Nash equilibrium consisting of zero contribitions by every subject. In experiments, a substantial proportion of subjects do indeed contribute zero. However, a sizeable proportion of subjects also contribute positively, with much variation in contributions both between and within subjects, and the objective of most experiments is to investigate the motivations behind such positive contributions.

In this part of the course, we develop a model which allows for a variety of motivations. It uses data from a real experiment (Bardsley 2000), whose design is tailored towards separate identification of these various motivations. The key feature of the experimental design is that subjects take turns to contribute, and each observes the contributions made by subjects placed earlier in the sequence. Also, the task is repeated, so the resulting data is panel data.

A number of econometric issues need to be addressed. Most importantly, it is clear from previous literature that there are different types of agent in the population, each with different contribution motives. Hence, a mixture model, of the type used earlier in other contexts, is appropriate for separating the various subject types.

The mixture model that we develop assumes three types. A "reciprocator" is a subject who contributes more when contributions by others placed earlier in the sequence are higher. We capture reciprocity by allowing the (reciprocator) subject's contribution to depend on the median of previous contributions within the sequence. A "strategist" is a subject who is selfish, but is willing to make positive contributions in anticipation of reciprocity by others placed later in the sequence. Since as the sequence progresses there are less subjects left to play, there should be less incentive for strategists to contribute the later a subject's position within the game. For example, for a strategist in last position, there is definitely no contribution motive. Hence for strategists, there should be an inverse relationship between contributions and subjects' position in the sequence. We therefore define "strategists" as agents whose contribution depends only on their position in the sequence. In last position we would expect them to contribute nothing. Finally, a "free-rider" is a subject who displays a tendency to contribute zero regardless of the behaviour of other subjects' or of their position in the sequence.

Application of the mixture framework in this setting is more complicated than in previous examples, because of the panel structure of the data. This is a further econometric issue to be addressed.

Another potential influence on a subject's contribution is task number. One almost universal finding in public goods games is a downward trend in contributions as the game is repeated. Standard explanations are in terms of a learning process. Subjects are learning, either about the game's incentive structure (i.e. learning to be rational), or about others' behaviour (*social* learning). A novel feature of this experimental design is that, for reasons to be explained in the next sub-section, it removes the effect of social learning; any decay in contributions over the course of the experiment will be attributable exclusively to learning about the incentive structure.

A final econometric issue is censoring. Contributions constitute a doubly censored dependent variable, since the lowest possible contribution is zero, and the highest possible contribution is the amount of the endowment. A two-limit Tobit model (Nelson 1976) is therefore required in order to obtain consistent estimates of the effects of experimental variables.

5.5.2 Experiment

We will use data from the experiment of Bardsley (2000). There are 98 subjects, divided into groups of size 7. Each subject performs 20 tasks.

The experimental design has a number of distinctive features. Firstly, within a single game, subjects take turns to contribute, and each subject observes the sequence of previous contributions. There are two reasons why this is important. Firstly, given that it is possible for a subject to observe previous contributions of others, it is possible for us to be able to assess the extent to which their contribution is driven by the previous contributions of others. That is, it becomes possible to test for reciprocity. Secondly, given that a subject are aware of their own position in the ordering, they will clearly know how many subjects are contributing after them, and they will therefore be in a position to assess the benefits from "strategic" contribution. By investigating the effect of position in the ordering on contribution, it therefore becomes possible for us to test for strategic behaviour.

Another distinctive feature of this experiment is the use of the "Conditional Information Lottery" (CIL) [9] as an incentive mechanism. In a CIL, the game played by the subject is camouflaged amongst a set of 19 (in this case) controlled fictional tasks. Conditional on a task's being the real one, the task information describes the real game (so "others' behaviour" is as shown). Subjects are told beforehand that only one task is a real game, that in the remainder, "others' behaviour" is simply an artefact of the design, and that only the real task is to be paid out. Subjects do not know which task is the real task, and it is therefore reasonable to suppose that they treat each task as if it is the real task.

The CIL is similar to the Random Lottery Incentive (RLI) mechanism, in the sense that only one from a set of tasks is for real and subjects do not know which is the real task until the end of the experiment. Unlike the RLI, however, the experimenter knows which task is for real from the start.

The main benefit from using the CIL is that it removes the effect of social learning. This is because subjects are aware that only one game is actually being played with the other subjects in the group. Given that a subject is temporarily assuming that the game currently being played is the real game, it is logical for that subject also to assume that all previous tasks were fictional, so anything learnt from those previous tasks cannot logically be about the behaviour of other subjects in the group. For this reason any decay in contributions over the course of the experiment will be attributable exclusively to learning about the incentive structure.

[9] see Bardsley (2000) for a full explanation of the CIL.

5.5.3 The data

The data set of Bardsley (2000) is contained in the file **bardsley**.

As noted previously, 98 subjects were observed over 20 tasks. This is clearly panel data, and the best way to present panel data is using the `xtline` command, preceded by `xtset`:

```
xtset i t
xtline y
```

The result is shown in Figure 20. This clearly shows the extent of between-subject heterogeneity. `Do similar graphs on ALL panel data!`

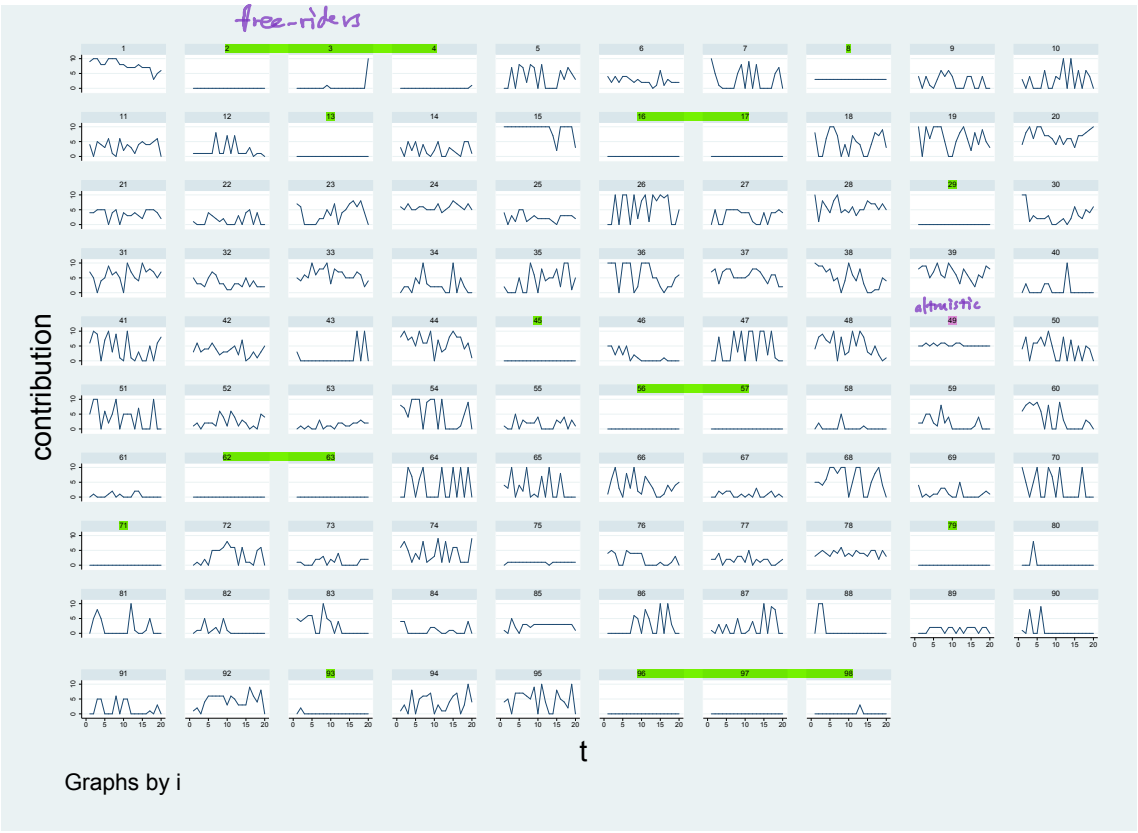


Figure 20: Time-series plots of contribution separately by subject

It is also revealing to examine the pooled distribution of contributions. A histogram of this variable is shown in Figure 21. The histogram clearly reveals censoring at zero, and to a lesser extent, censoring at the upper limit, 10. The overall mean contribution was 2.711, compared with a median of 1.0, this difference confirming the clear positive skew evident in the histogram.

In order to give a feel for the extent of between-subject variation, Figure 22 shows the distribution of the number of zero contributions made by each subject. Apart from establishing the wide variation in behaviour, Figure 22 is useful in providing a rough estimate of the number of free riders in the sample. Remembering the (strict) definition of a free-rider as a subject contributing zero on *every* occasion, we see that 14 of the 98 subjects (14.3%) satisfy this definition. However, the definition of

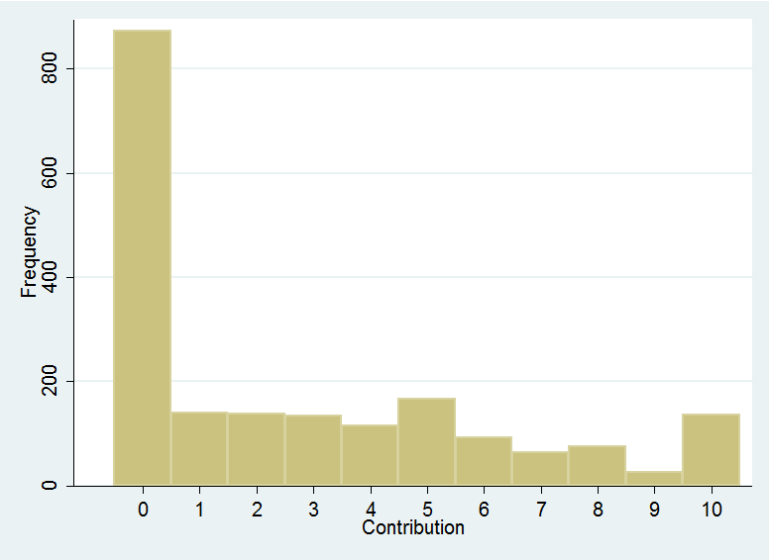


Figure 21: Distribution of contributions in Bardsley's experiment

free riders which we incorporate in estimation is less rigid than this. We include a tremble parameter which allows a small probability of subjects setting their contribution completely at random on any task. This means that a genuine free rider may be observed making positive contributions on a small number of occasions. Fig 22 is useful because the clearly discernible cluster of 24 subjects (24.5%) who contribute zero on at least 16 out of 20 occasions, may reasonably be tentatively identified as free-riders who are subject to a tremble.

It is also useful to use graphical methods to investigate the nature of the effects of the various determinants of contribution. For this purpose, we present three scatter plots with lowess smoothers in Fig 23. Since the effects of these variables cannot apply to the behaviour of free-riders, contributions of the 24 subjects identified loosely as free-riders in the context of Fig 22 above are excluded from Fig 23. The scatters themselves are clearly uninformative since the vast majority of the possible combinations of MED and contribution are represented in the plots. For this reason, we include "jitter" in the scatterplots to make it possible at least to see which locations contain the most points. We also include Lowess smoothers, which plot the estimated conditional mean of contribution.

The Lowess smoothers in Fig 23 reveal that all three variables appear to have an impact on contribution. Moreover, the direction of each effect is in accordance with theoretical predictions: the median of previous contributions has a positive effect, as predicted by reciprocity theory; order in the sequence has a negative effect, implying strategic behaviour; task number has a negative effect, implying a process of learning about the game. We also see that the three effects are monotonic and roughly linear. These observations are useful in guiding the specification of the parametric model developed in the next section.

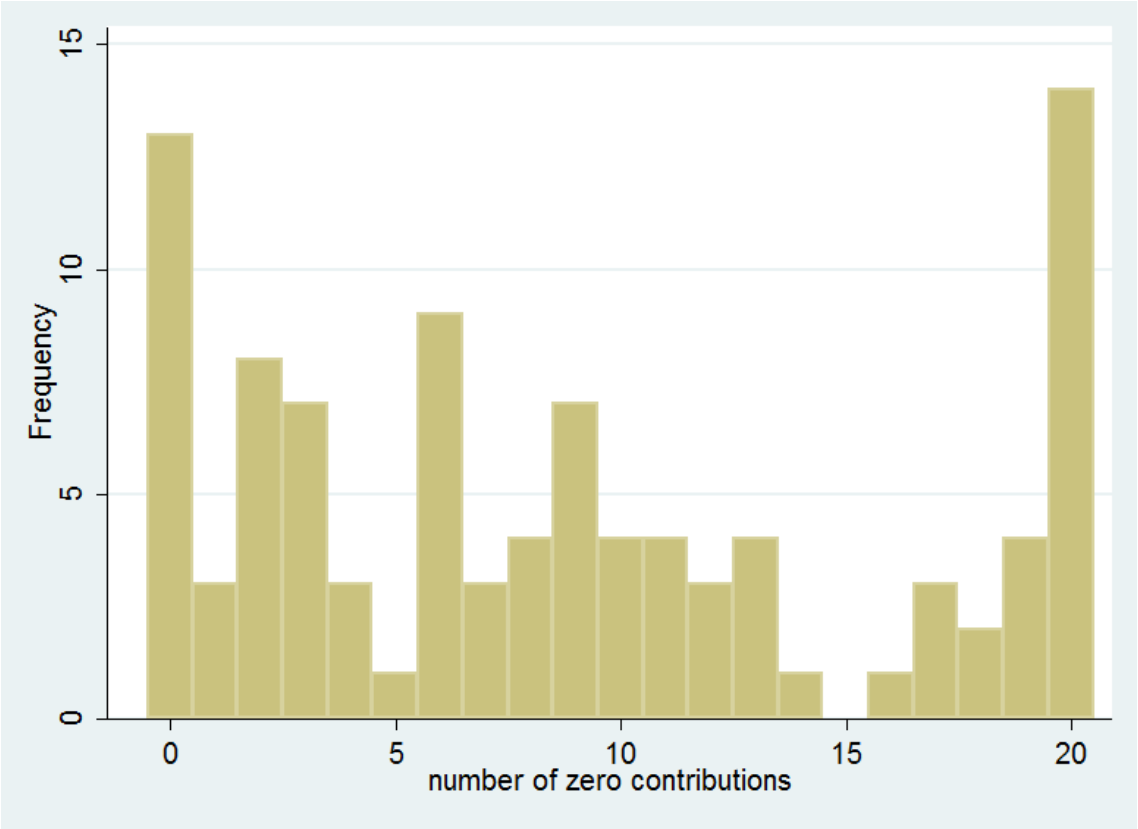


Figure 22: A histogram of the number of zero contributions by each subject

5.5.4 The Finite Mixture 2-Limit Tobit Model with tremble

The econometric analysis used here is similar to that of Bardsley & Moffatt (2007).

We assume that there are n subjects, each of whom has been observed over T tasks. Let y_{it} be the observed contribution by subject i in task t. The variable y_{it} has a lower limit of 0 and an upper limit of 10. The two-limit tobit model (Nelson 1976), with limits 0 and 10, is therefore appropriate. To adopt conventional terminology in limited dependent variable modelling, we refer to zero contributions as being in "regime 1", contributions greater than 0 but less than 10 in "regime 2", and contributions of 10 in "regime 3".

The underlying desired contribution is y^*_{it} and this will be assumed to depend linearly on a set of explanatory variables. However, as explained in Sect 5.5.1, we are assuming that each subject is one of three types: reciprocator (rec), strategist (str) and free-rider (fr), and the determination of y^*_{it} for a given subject depends crucially on which type that subject is. An important feature of the model (and finite mixture models in general) is that there is no switching between types: given that a particular subject is of a given type, the subject is of that type for every task undertaken.

For reciprocators and strategists, we specify the following latent model for the desired contribution (free-riders are treated separately since no explanatory variables are relevant to their contribution):

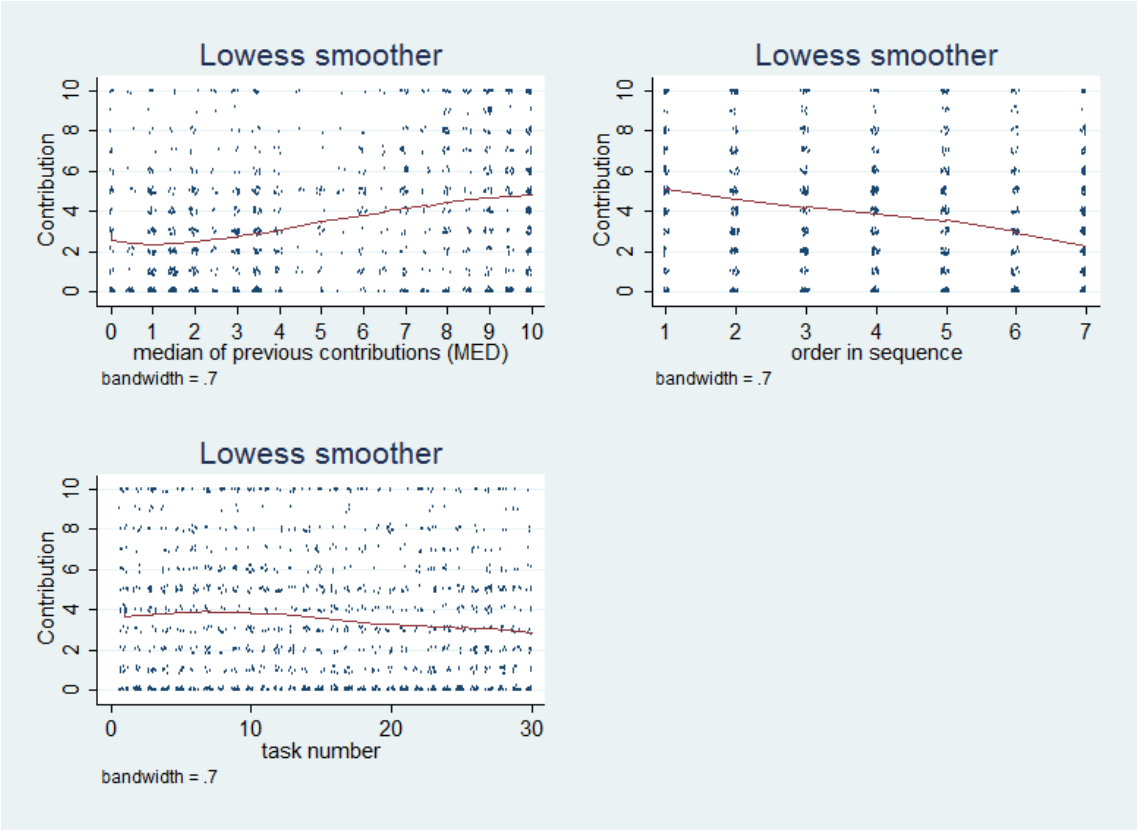


Figure 23: Jittered scatter plots and Lowess smoothers of contribution against (i) median of previous contributions (ii) order in sequence (iii) task number. Free-riders excluded.

$$\begin{aligned} \text{reciprocator:} \quad & y^*_{it} = \beta_{10} + \beta_{11} MED_{it} + \beta_{13}(TSK_{it} - 1) + \epsilon_{it,rec} \\ \text{strategist:} \quad & y^*_{it} = \beta_{20} + \beta_{22}(ORD_{it} - 1) + \beta_{23}(TSK_{it} - 1) + \epsilon_{it,str} \\ & i = 1, ..., n \quad t = 1, ..., T \\ & \epsilon_{it,rec} \sim N(0, \sigma_1^2) \qquad \epsilon_{it,str} \sim N(0, \sigma_2^2) \end{aligned} \tag{77}$$

where ORD_{it} is subject i's position in the group for the tth task solved, MED_{it} is the median of previous contributions by other subjects in the group , and TSK_{it} is the task number [10]. Reciprocity implies $\beta_{11} > 0$, while strategic behaviour implies $\beta_{22} < 0$. The parameters β_{13} and β_{23} represent learning by reciprocators and strategists respectively, and are expected to be negative. The reason for subtracting one from TSK and ORD is to ensure that the intercept in each equation has a convenient interpretation: expected contribution for a subject in first position in the first task.

When $ORD = 1$, MED is clearly not defined, but (following Bardsley & Moffatt (2007, see their Table II, note 4)) will be set to 8.00 for the purpose of estimation; this value is obtained by a trial and error process, being that which maximises the log-likelihood. It could usefully be interpreted as subjects' *a priori* expectation of

[10] TSK is not the same as t, since some of the tasks are part of a separate experiment. While t goes from 1 to 20, the range of TSK is from 1 to 30.

others' contributions, formed before the start of the experiment. The relatively high value embodies the idea that reciprocators start the game with an optimistic outlook regarding the generosity of other players.

The relationship between *desired* contribution y^*_{it} and *actual* contribution y_{it} is specified by the following censoring rules:

For reciprocators and strategists:

$$y_{it} = \begin{cases} 0 & \text{if} \quad y^*_{it} \leq 0 \\ y^*_{it} & \text{if} \quad 0 < y^*_{it} < 10 \\ 10 & \text{if} \quad y^*_{it} \geq 10 \end{cases} \tag{78a}$$

For free riders:

$$y_{it} = 0 \quad \forall t \tag{78b}$$

As mentioned in sect 5.5.3, we also introduce a "tremble parameter", ω (see Moffatt & Peters (2001)). On any single response, with probability ω a subject loses concentration and chooses randomly between the eleven possible contributions. One purpose of this parameter is to relax the rigid segregation rules between the three subject types. If, for example, a subject contributes zero on every occasion except one, we wish to assign a positive probability to this subject being a free-rider who lost concentration on one occasion. The presence of the tremble parameter allows this.

Loomes et al. (2002), in their econometric models of risky choice, include a tremble parameter which decays in magnitude in the course of the experiment, to allow for a learning process: subjects are less likely to make "random" choices when they have more experience. A similar strategy is adopted here. We specify:

$$\omega_{it} = \omega_0 \exp\left[\omega_1 (TSK_{it} - 1)\right] \tag{79}$$

There are now two parameters associated with the tremble: ω_0 represents the tremble probability at the start of the experiment, while ω_1 represents the rate of decay. We expect ω_1 to take a negative value, and the larger it is in magnitude, the faster the implied decay.

For each regime and each subject type, we have the following likelihood contributions for a single response, where $\Phi(.)$ and $\phi(.)$ are the standard normal c.d.f. and p.d.f. respectively:

Regime 1 $(y = 0)$:

$$\begin{aligned} P\left(y_{it} = 0 | i = \text{rec}\right) &= (1 - \omega_{it})\Phi\left(-\frac{\beta_{10} + \beta_{11} MED_{it} + \beta_{13}(TSK_{it} - 1)}{\sigma_1}\right) + \frac{\omega_{it}}{11} \\ P\left(y_{it} = 0 | i = \text{str}\right) &= (1 - \omega_{it})\Phi\left(-\frac{\beta_{20} + \beta_{22}(ORD_{it} - 1) + \beta_{23}(TSK_{it} - 1)}{\sigma_2}\right) + \frac{\omega_{it}}{11} \\ P\left(y_{it} = 0 | i = \text{fr}\right) &= 1 - \frac{10\omega_{it}}{11} \end{aligned} \tag{80a}$$

Regime 2 $(0 < y < 10)$:

$$f\left(y_{it}|i=\text{rec}\right) = (1-\omega_{it})\frac{1}{\sigma_1}\phi\left(\frac{y_{it}-\beta_{10}-\beta_{11}MED_{it}-\beta_{13}(TSK_{it}-1)}{\sigma_1}\right)+\frac{\omega_{it}}{11}$$
$$f\left(y_{it}|i=\text{str}\right) = (1-\omega_{it})\frac{1}{\sigma_2}\phi\left(\frac{y_{it}-\beta_{20}-\beta_{22}(ORD_{it}-1)-\beta_{23}(TSK_{it}-1)}{\sigma_2}\right)+\frac{\omega_{it}}{11}$$
$$f\left(y_{it}|i=\text{fr}\right) = \frac{\omega_{it}}{11} \tag{80b}$$

Regime 3 $(y = 10)$:

$$P\left(y_{it}=10|i=\text{rec}\right) = (1-\omega_{it})\left[1-\Phi\left(\frac{10-\beta_{10}-\beta_{11}MED_{it}-\beta_{13}(TSK_{it}-1)}{\sigma_1}\right)\right]+\frac{\omega_{it}}{11}$$
$$P\left(y_{it}=10|i=\text{str}\right) = (1-\omega_{it})\left[1-\Phi\left(\frac{10-\beta_{20}-\beta_{22}(ORD_{it}-1)-\beta_{23}(TSK_{it}-1)}{\sigma_2}\right)\right]+\frac{\omega_{it}}{11}$$
$$P\left(y_{it}=10|i=\text{fr}\right) = \frac{\omega_{it}}{11} \tag{80c}$$

The manner in which the tremble parameter appears in (80) may require explanation. When a tremble occurs, each of the 11 outcomes, 0-10, is each equally likely, hence the term $\omega_{it}/11$ appearing in nearly every equation. In regime 2, what is required is a density, not a probability, so we imagine that when a tremble occurs contributions are realisations from a continuous uniform distribution on $(-0.5, 10.5)$, whence the density associated with any particular realisation is $\omega_{it}/11$.

It is the existence of the three distinct types of subject that leads to a finite mixture model. We introduce three "mixing proportions", p_{rec}, p_{str} and p_{fr}, which represent the proportion of the population who are reciprocators, strategists and free-riders respectively. Since these three parameters sum to unity, only two are estimated.

The Likelihood contribution for subject i is:

$$\begin{aligned}L_i = p_{\text{rec}}\prod_{t=1}^{T} P(y_{it}=0|\text{rec})^{I_{y_{it}=0}} f(y_{it}|\text{rec})^{I_{0<y_{it}<10}} P(y_{it}=10|\text{rec})^{I_{y_{it}=10}} \\ +p_{\text{str}}\prod_{t=1}^{T} P(y_{it}=0|\text{str})^{I_{y_{it}=0}} f(y_{it}|\text{str})^{I_{0<y_{it}<10}} P(y_{it}=10|\text{str})^{I_{y_{it}=10}} \\ +p_{\text{fr}}\prod_{t=1}^{T} P(y_{it}=0|\text{fr})^{I_{y_{it}=0}} f(y_{it}|\text{fr})^{I_{0<y_{it}<10}} P(y_{it}=10|\text{fr})^{I_{y_{it}=10}}\end{aligned} \tag{81}$$

where $I_{(.)}$ is the indicator function (taking the value 1 if the subscripted expression is true, 0 otherwise), and the nine conditional probabilities/densities are specified in (80) above.

The sample log-likelihood is then:

$$LogL = \sum_{i=1}^{n}\log\left(L_i\right) \tag{82}$$

$LogL$ is maximised to obtain MLEs of the eight parameters appearing in (80), and

in addition the two tremble parameters and two of the three mixing proportions. The total number of estimated parameters in the full model is 12. This model may be described as the "finite-mixture 2-limit tobit model with tremble".

5.5.5 Program

As mentioned, the panel structure of this data set is a complicating feature. Each subject is observed 20 times. When computing the likelihood contribution for a given subject, we require the joint probability of all 20 of the decisions made by that subject. In practical terms, this means that we require a different likelihood evaluator in the ML program from that used in previous examples.

There are a number of different likelihood evaluators in STATA. The one that was used earlier was *lf* (linear form). A feature of the log-likelihood function defined in (81) and (82) is that it does not satisfy the linear form restrictions, and therefore *lf* cannot be used. This is because the likelihood contributions that need to be summed in order to obtain the sample log-likelihood are not each derived from the information in a single row of the data, but are instead derived from the entire block of rows corresponding to a given subject. There is only one likelihood contribution for each such block of rows. Because of this, the *d-family* evaluators are required in place of the *lf* evaluator. The simplest of these is the *d0* evaluator, which simply requires the log-likelihood contributions to be evaluated. This is the one that will be used here. The *d1* and *d2* evaluators require analytical derivatives of the log-likelihood to be programmed as well as the function evaluation.

The STATA code is presented below. Table 5 gives the names in the code corresponding to each of the components in the construction of $LogL$ above. One section of the code which may require explanation is:

```
by i: replace ‘pp1’=exp(sum(ln(max(‘p1’,1e-12))))
by i: replace ‘pp2’=exp(sum(ln(max(‘p2’,1e-12))))
by i: replace ‘pp3’=exp(sum(ln(max(‘p3’,1e-12))))
```

This essentially takes the product of the probabilities contained in (in the first case) p1, over all T observations for subject i. The reason why we apply the three functions exp(sum(ln(.))) is simply because, although STATA has a "sum" function (which takes the sum of a variable over observations), it does not have a a "product" function. Hence we evaluate the required product by exploiting the identity:

$$\prod_t p_t \equiv \exp\left(\sum_t \ln p_t\right) \tag{83}$$

The reason why we take the log of $max(p1, 1e-12)$, rather than simply $p1$, is to prevent the numerical problems that would arise if the probabilities were ever extremely close to zero.

Note that the mixing proportion for the third type (p_3) is deduced from p_1 and p_2 using the delta method. Note also that the final section of the code generates posterior type probabilities. This will be discussed in section 5.5.7.

The annotated code is as follows.

Component of $LogL$	STATA name
β_{10}, β_{11}, β_{13}	theta1
β_{20}, β_{22}, β_{23}	theta2
σ_1, σ_2	sig1, sig2
ω_0, ω_1, ω	w0, w1, w
p_{rec}, p_{str}, p_{fr}	p_rec, p_str, p_fr
$P(y=0\|\text{rec})$, $P(y=0\|\text{str})$, $P(y=0\|\text{fr})$	p1_1, p2_1, p3_1
$f(y\|\text{rec})$, $f(y\|\text{str})$, $f(y\|\text{fr})$; $0<y<10$	p1_2, p2_2, p3_3
$P(y=10\|\text{rec})$, $P(y=10\|\text{str})$, $P(y=10\|\text{fr})$	p1_3, p2_3, p3_3
$P(y_{it}=0\|\text{rec})^{I_{y_{it}=0}} f(y_{it}\|\text{rec})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{rec})^{I_{y_{it}=10}}$	p1
$P(y_{it}=0\|\text{str})^{I_{y_{it}=0}} f(y_{it}\|\text{str})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{str})^{I_{y_{it}=10}}$	p2
$P(y_{it}=0\|\text{fr})^{I_{y_{it}=0}} f(y_{it}\|\text{fr})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{fr})^{I_{y_{it}=10}}$	p3
$\prod_{t=1}^{T} P(y_{it}=0\|\text{rec})^{I_{y_{it}=0}} f(y_{it}\|\text{rec})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{rec})^{I_{y_{it}=10}}$	pp1
$\prod_{t=1}^{T} P(y_{it}=0\|\text{str})^{I_{y_{it}=0}} f(y_{it}\|\text{str})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{str})^{I_{y_{it}=10}}$	pp2
$\prod_{t=1}^{T} P(y_{it}=0\|\text{fr})^{I_{y_{it}=0}} f(y_{it}\|\text{fr})^{I_{0<y_{it}<10}} P(y_{it}=10\|\text{fr})^{I_{y_{it}=10}}$	pp3
L_i	pp
$LogL = \sum_{i=1}^{n} \log(L_i)$	lnpp
$P(i=\text{rec}\|y_{i1},...,y_{iT})$; $P(i=\text{str}\|y_{i1},...,y_{iT})$; $P(i=\text{fr}\|y_{i1},...,y_{iT})$	postp1; postp2; postp3

Table 5: Components of $LogL$ and corresponding STATA names.

```
* ESTIMATION OF MIXTURE MODEL FOR BARDSLEY DATA

prog drop _all

* LIKELIHOOD EVALUATION PROGRAM STARTS HERE:

program define pg_mixture
args todo b lnpp
tempvar p1_1 p2_1 p3_1 p1_2 p2_2 p3_2 p1_3 p2_3 p3_3 p1 p2 p3 pp1 pp2 pp3 pp w

tempname theta1 theta2 sig1 sig2 w0 w1 p_rec p_str

* ASSIGN PARAMETER NAMES TO THE ELEMENTS OF THE PARAMETER VECTOR b:

mleval `theta1' = `b', eq(1)
mleval `theta2' = `b', eq(2)
mleval `sig1' = `b', eq(3) scalar
mleval `sig2'=`b', eq(4) scalar
mleval `w0'=`b', eq(5) scalar
mleval `w1'=`b', eq(6) scalar
mleval `p_rec'=`b', eq(7) scalar
mleval `p_str'=`b', eq(8) scalar

quietly{

* INITIALISE THE p* VARIABLES WITH MISSING VALUES:

gen double `p1_1'=.
gen double `p2_1'=.
gen double `p3_1'=.
gen double `p1_2'=.
gen double `p2_2'=.
gen double `p3_2'=.
gen double `p1_3'=.
gen double `p2_3'=.
gen double `p3_3'=.

gen double `p1'=.
gen double `p2'=.
gen double `p3'=.

gen double `pp1'=.
gen double `pp2'=.
gen double `pp3'=.
```

```
gen double ‘pp’=.

* GENERATE THE TREMBLE PROBABILITY:

gen double ‘w’=‘w0’*exp(‘w1’*tsk_1)

* COMPUTE TYPE-CONDITIONAL DENSITIES UNDER REGIME 1:

replace ‘p1_1’=(1-‘w’)*normal(-‘theta1’/‘sig1’)+‘w’/11
replace ‘p2_1’=(1-‘w’)*normal(-‘theta2’/‘sig2’)+‘w’/11
replace ‘p3_1’=1-(10/11)*‘w’

* COMPUTE TYPE-CONDITIONAL DENSITIES UNDER REGIME 2:

replace ‘p1_2’=(1-‘w’)*(1/‘sig1’)*normalden((y-‘theta1’)/‘sig1’)+‘w’/11
replace ‘p2_2’=(1-‘w’)*(1/‘sig2’)*normalden((y-‘theta2’)/‘sig2’)+‘w’/11
replace ‘p3_2’=‘w’/11

* COMPUTE TYPE-CONDITIONAL DENSITIES UNDER REGIME 3:

replace ‘p1_3’=(1-‘w’)*(1-normal((10-‘theta1’)/‘sig1’))+‘w’/11
replace ‘p2_3’=(1-‘w’)*(1-normal((10-‘theta2’)/‘sig2’))+‘w’/11
replace ‘p3_3’=‘w’/11

* MATCH TYPE-CONDITIONAL DENSITIES TO ACTUAL REGIMES (d IS REGIME):

replace ‘p1’ = (d==1)*‘p1_1’+(d==2)*‘p1_2’+(d==3)*‘p1_3’
replace ‘p2’ = (d==1)*‘p2_1’+(d==2)*‘p2_2’+(d==3)*‘p2_3’
replace ‘p3’ = (d==1)*‘p3_1’+(d==2)*‘p3_2’+(d==3)*‘p3_3’

* FIND PRODUCT OF TYPE-CONDITIONAL DENSITIES FOR EACH SUBJECT:

by i: replace ‘pp1’=exp(sum(ln(max(‘p1’,1e-12))))
by i: replace ‘pp2’=exp(sum(ln(max(‘p2’,1e-12))))
by i: replace ‘pp3’=exp(sum(ln(max(‘p3’,1e-12))))

* COMBINE TYPE-CONDITIONAL DENSITIES TO OBTAIN MARGINAL DENSITY FOR EACH SUBJECT
* (ONLY REQUIRED IN FINAL ROW FOR EACH SUBJECT):

replace ‘pp’=‘p_rec’*‘pp1’+‘p_str’*‘pp2’+(1-‘p_rec’-‘p_str’)*‘pp3’
replace ‘pp’=. if last~=1

* SPECIFY (LOG-LIKELIHOOD) FUNCTION WHOSE SUM OVER SUBJECTS IS TO BE MAXIMISED

mlsum ‘lnpp’=ln(‘pp’) if last==1

* GENERATE POSTERIOR TYPE PROBABILITIES, AND MAKE THESE AVAILABLE OUTSIDE THE PROGRAM

replace postp1=‘p_rec’*‘pp1’/‘pp’
replace postp2=‘p_str’*‘pp2’/‘pp’
replace postp3=(1-‘p_rec’-‘p_str’)*‘pp3’/‘pp’

putmata postp1, replace
putmata postp2, replace
putmata postp3, replace

}

end

* END OF LOG-LIKELIHOOD EVALUATION PROGRAM

clear
set more off

* READ DATA

use "bardsley"

by i: gen last=_n==_N

gen int d=1
replace d=2 if y>0
replace d=3 if y==10

gen double ord_1=ord-1
```

```
gen double tsk_1=tsk-1

* SET MEDIAN OF PREVIOUS CONTRIBUTIONS TO 8 FOR SUBJECTS IN FIRST POSITION:

replace med=8 if ord==1

*  SPECIFY EXPLANATORY-VARIABLE LISTS FOR RECIPROCATOR (LIST1)
*  AND STRATEGIST (LIST2) EQUATIONS:

local list1 "med tsk_1"
local list2 "ord_1 tsk_1"

* INITIALISE VARIABLES TO BE USED FOR POSTERIOR TYPE PROBABILITIES:

gen postp1=.
gen postp2=.
gen postp3=.

* SPECIFY STARTING VALUES:

mat start=(0.57,-0.10,6.1,-0.93,-0.05,5.2,3.3,3.7,0.11,-0.05,0.26,0.49)

* SPECIFY LIKELIHOOD EVALUATOR, PROGRAM, AND PARAMETER NAMES:

ml model d0 pg_mixture (=‘list1’) (=‘list2’) /sig1 /sig2 /w0  /w1  /p1 /p2
ml init start, copy

* USE ML COMMAND TO MAXIMISE LOG-LIKELIHOOD, AND STORE RESULTS AS "WITH_TREMBLE":

ml max, trace search(norescale)
est store with_tremble

* COMPUTE THIRD MIXING PROPORTION USING DELTA METHOD:

nlcom p3: 1-[p1]_b[_cons]-[p2]_b[_cons]

* EXTRACT POSTERIOR TYPE PROBABILITIES AND PLOT THEM AGAINST
*  NUMBER OF ZERO CONTRIBUTIONS:

drop postp1 postp2 postp3

getmata postp1
getmata postp2
getmata postp3

label variable postp1 "rec"
label variable postp2 "str"
label variable postp3 "fr"

by i: gen n_zero=sum(y==0)

scatter postp1 postp2 postp3 n_zero if last==1, title("with tremble") ///
ytitle("posterior probability") msymbol(x Dh Sh) jitter(3) saving(with, replace)

* ESTIMATE MODEL WITHOUT TREMBLE, AND STORE RESULTS AS "WITHOUT_TREMBLE":

constraint 1 [w0]_b[_cons]=0.00
constraint 2 [w1]_b[_cons]=0.00

ml model d0 pg_mixture (=‘list1’) (=‘list2’) ///
/sig1 /sig2 /w0  /w1  /p1 /p2, constraints(1 2)

ml init start, copy
ml max, trace search(norescale)
est store without_tremble

nlcom p3: 1-[p1]_b[_cons]-[p2]_b[_cons]

* EXTRACT AND PLOT POSTERIOR TYPE PROBABILITIES FOR MODEL WITHOUT TREMBLE:

drop postp1 postp2 postp3

getmata postp1
getmata postp2
getmata postp3
```

```
label variable postp1 "rec"
label variable postp2 "str"
label variable postp3 "fr"

scatter postp1 postp2 postp3 n_zero if last==1, title("without tremble") ///
ytitle("posterior probability") msymbol(x Dh Sh) jitter(3) saving(without, replace)

* CARRY OUT LIKELIHOOD RATIO TEST FOR PRESENCE OF TREMBLE:

lrtest with_tremble without_tremble

* COMBINE THE TWO POSTERIOR PROBABILITY PLOTS

gr combine with.gph without.gph
```

The model is estimated twice, first with all parameters unconstrained, and second with the two tremble parameters constrained to zero. Note that all that is required for this is to define the two constraints using the "constraint" command, and then to include the constraints(.) option with the "ml" command.

The STATA output from the first estimation (the model with tremble) is as follows:

```
                                                Number of obs   =       1960
                                                Wald chi2(2)    =     108.07
Log likelihood = -3267.6884                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
         med |    .598677   .0611812     9.79   0.000     .4787641    .7185899
       tsk_1 |  -.0961739   .0202229    -4.76   0.000      -.13581   -.0565379
       _cons |   4.004374   .4541832     8.82   0.000     3.114192    4.894557
-------------+----------------------------------------------------------------
eq2          |
       ord_1 |  -.9644643   .0823741   -11.71   0.000    -1.125915    -.803014
       tsk_1 |  -.0516766    .017189    -3.01   0.003    -.0853664   -.0179867
       _cons |   5.299353   .3828498    13.84   0.000     4.548981    6.049724
-------------+----------------------------------------------------------------
sig1         |
       _cons |   3.442241   .1674649    20.56   0.000     3.114016    3.770466
-------------+----------------------------------------------------------------
sig2         |
       _cons |   3.705603   .1611296    23.00   0.000     3.389794    4.021411
-------------+----------------------------------------------------------------
w0           |
       _cons |    .104174   .0321192     3.24   0.001     .0412216    .1671265
-------------+----------------------------------------------------------------
w1           |
       _cons |  -.0492262   .0218191    -2.26   0.024    -.0919909   -.0064614
-------------+----------------------------------------------------------------
p1           |
       _cons |   .2710853    .048467     5.59   0.000     .1760918    .3660788
-------------+----------------------------------------------------------------
p2           |
       _cons |   .4832814   .0538021     8.98   0.000     .3778311    .5887316
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          p3 |   .2456333   .0436144     5.63   0.000     .1601506     .331116
------------------------------------------------------------------------------
```

As recommended by Moffatt & Peters (2001), the Likelihood Ratio test has been used to test for the presence of a tremble. Results from the likelihood ratio test comparing the above model with the tremble-free model are as follows. The p-value of 0.0000 represents overwhelming evidence of the presence of a tremble.

```
. lrtest with_tremble without_tremble

Likelihood-ratio test                                 LR chi2(2)  =    149.89
(Assumption: without_trem~e nested in with_tremble)   Prob > chi2 =    0.0000
```

Results from both models are presented and discussed in the next section.

5.5.6 Results

The parameter estimates from the finite mixture 2-limit tobit model, with and without tremble, are shown in table 6. The first column contains results from the full model. The second column shows the results of the model with no tremble. As noted at the end of the last section, the likelihood ratio test provided overwhelming evidence of the presence of a tremble in this data set, implying that the full model is superior. The importance of including the tremble is also clearly seen by observing how different the estimates are, in particular the mixing proportions, when the tremble is absent. For example, in the absence of a tremble, the proportion of free riders is estimated to be 0.143 which is exactly equal to the proportion of subjects who donated zero on every occasion (see section 5.5.3); as mentioned previously, the presence of the tremble allows the set of free-riders to include subjects who donated zero on *nearly* all occasions, and accordingly, the estimate of the proportion of free-riders rises to 0.246, which is remarkably close to the proportion of subjects non-parametrically identified as free riders in Section 5.5.3, on the grounds that they contributed zero in at least 16 out of 20 tasks. This estimate of 24.6% is also in close agreement with previous estimates appearing elsewhere in the literature.
The two equations set out in (77) are estimated as:

$$\text{Reciprocators}: \quad E(y^*|MED, TSK) = 4.004 + 0.599MED - 0.096(TSK - 1) \tag{84a}$$

$$\text{Strategists}: \quad E(y^*|ORD, TSK) = 5.299 - 0.964(ORD - 1) - 0.052(TSK - 1) \tag{84b}$$

As seen in table 6, all coefficients are strongly significant. For reciprocators, as expected, the median of previous contributions has a significantly positive effect on the current contribution: if all previous contributions were raised by one unit, we would expect the current contribution to rise by around three-fifths of one unit, but by significantly less than one whole unit. This result is consistent with the biased reciprocity observed in Fischbacher et al. (2001) (biased in the sense that subjects, although influenced positively by the contributions of others, tend to donate less than the levels contributed by others). This bias, which we have identified using a one-shot sequential play game, may be partly responsible for the usually-observed decay of contributions in the more usual environment of simultaneous play, repeated game experiments.

For strategists, again as expected, the effect of the subjects' order in the sequence is negative. In particular, the "expected" contribution of a strategist first-mover (in task 1) is 5.3, while the same strategist in last position ($ORD = 7$) would be expected to contribute zero - a highly reassuring result, since as we noted earlier there is no selfish contribution motive for a subject in last position. In contrast, reciprocators are never expected to contribute zero.

The effect of TSK is significantly negative for both types, simply implying a diminution of contributions with experience. If this is interpreted as the effect of learning about the incentive structure of the game, it seems that reciprocators learn about such matters somewhat faster than strategists.

The tremble probability is 0.104 at the start of the experiment (task 1), but, in accordance with the significant negative estimate of ω_1, decays to 0.041 by the end (task 20). This dramatic decay of the tremble amounts to further evidence of learning (Moffatt & Peters 2001, Loomes et al. 2002).

Turning to the estimates of the mixing proportions, we see that very close to 25% of the population are free riders; around 25% are reciprocators; the remaining 50% are strategists.

	full model	no tremble
Reciprocators		
constant	4.004(0.454)	3.166(0.358)
MED	0.599(0.061)	0.490(0.045)
TSK-1	$-0.096(0.020)$	$-0.061(0.015)$
σ_1	3.442(0.167)	3.577(0.126)
Strategists		
constant	5.299(0.382)	4.493(0.518)
ORD-1	-0.964(0.082)	-1.128(0.102)
TSK-1	-0.052(0.017)	-0.080(0.023)
σ_2	3.706(0.161)	5.104(0.253)
Tremble		
ω_0	0.104(0.032)	-
ω_1	-0.049(0.022)	-
Mixing proportions		
p_{rec}	0.271(0.048)	0.382(0.051)
p_{str}	0.483(0.054)	0.472(0.053)
p_{fr}	0.246(0.044)	0.143(0.035)
n	98	98
T	20	20
k	12	10
LogL	-3267.69	-3342.63
AIC	3.35	3.42

Table 6: Maximum Likelihood estimates from mixture model applied to Bardsley (2000)'s data, with and without tremble. Asymptotic standard errors in parentheses. The estimate and standard error of p_{fr} is deduced from the estimates of p_{rec} and p_{str} using the delta method. When ORD=1, MED is set to 8 for the purpose of estimation. AIC is Akaike's Information Criterion, defined as $2(-LogL + k)/(nT)$, where k is the number of parameters in the model. The preferred model is the one with the lower AIC.

5.5.7 Posterior Type Probabilities

The three posterior type probabilities are given by:

$$P(i = \text{rec}|y_{i1}, ..., y_{iT}) = \frac{p_{\text{rec}} \prod_{t=1}^{T} P(y_{it} = 0|\text{rec})^{I_{y_{it}=0}} f(y_{it}|\text{rec})^{I_{0<y_{it}<10}} P(y_{it} = 10|\text{rec})^{I_{y_{it}=10}}}{L_i}$$

$$P(i = \text{str}|y_{i1}, ..., y_{iT}) = \frac{p_{\text{str}} \prod_{t=1}^{T} P(y_{it} = 0|\text{str})^{I_{y_{it}=0}} f(y_{it}|\text{str})^{I_{0<y_{it}<10}} P(y_{it} = 10|\text{str})^{I_{y_{it}=10}}}{L_i}$$

$$P(i = \text{fr}|y_{i1}, ..., y_{iT}) = \frac{p_{\text{fr}} \prod_{t=1}^{T} P(y_{it} = 0|\text{fr})^{I_{y_{it}=0}} f(y_{it}|\text{fr})^{I_{0<y_{it}<10}} P(y_{it} = 10|\text{fr})^{I_{y_{it}=10}}}{L_i}$$

where L_i is the likelihood contribution for subject i, defined in (81). These posterior probabilities are computed (as postp1-postp3) at the end of the program.

In Figure 23, we plot the three posterior probabilities, obtained from estimation of both models, against the number of zero contributions made by the subject. Comparing the two plots, we see once again that the main difference between the two models is in the classification of subjects to the "free rider" type. For the tremble-free model (right-hand plot), only subjects contributing zero in all 20 tasks are classified as "free-rider". For the model with tremble however (left-hand plot), all subjects who contributed zero in 16 or more tasks are seen to be very likely to be free-riders. Further inspection of the left-hand plot reveals that subjects who contribute zero in a moderate number of tasks (6-14) tend to be strategists, while subjects who rarely contribute zero appear to be a mixture of strategists and reciprocators. Note finally that there few points in the scatter are far from zero or one on the vertical axis, indicating that it is only for a small number of subjects that the model is incapable of detecting type with confidence.

Figure 24: jittered scatter of posterior type probabilities against number of zero contributions from model with tremble (left-hand graph) and from model without tremble (right-hand graph)

References

Andreoni, J. & Miller, J. (2002), 'Giving according to GARP: An experimental test of the consistency of preferences for altruism', *Econometrica* **70**, 737–753.

Bardsley, N. (2000), 'Control without deception: Individual behaviour in free-riding experiments revisited', *Experimental Economics* **3**, 215–240.

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

Becker, G., DeGroot, M. & Marschak, J. (1964), 'Measuring utility by a single-response sequential method', *Behavioural Science* **9**, 226–232.

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

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

Eckel, C. & Grossman, P. J. (2001), 'Chivalry and solidarity in ultimatum games', *Economic Inquiry* **39**, 171–188.

Engelmann, D. & Strobel, M. (2004), 'Inequality aversion, efficiency, and maximin preferences in simple distribution experiments', *American economic review* pp. 857–869.

Fechner, G. (1860), *Elements of Psychophysics, Vol. 1*, New York: Holt, Rinehart and Winston.

Fehr, E. & Schmidt, K. M. (1999), 'A theory of fairness, competition and cooperation', *Quarterly Journal of Economics* **114**, 817–868.

Feiveson, A. H. et al. (2002), 'Power by simulation', *Stata J* **2**(2), 107–124.

Fischbacher, U., Gachter, S. & Fehr, E. (2001), 'Are people conditionally cooperative? evidence from a public goods experiment', *Economics Letters* **71**, 379–404.

Georg, S. J. (2009), 'Nonparametric testing of distributions - the Epps-Singleton two-sample test using the empirical characteristic function', *The Stata Journal* **9**, 454–465.

Harrison, G., Johnson, E., McInnes, M. & Rutstrom, E. E. (2005), 'Risk aversion and incentive effects: Comment', *American Economic Review* **95**, 900–904.

Holt, C. & Laury, S. K. (2002), 'Risk aversion and incentive effects', *American Economic Review* **92**, 1644–1655.

Isoni, A., Loomes, G. & Sugden, R. (2011), 'The willingness to pay willingness to accept gap, the" endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Comment', *American Economic Review* **101**(2), 991–1011.

Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1990), 'Experimental tests of the endowment effect and the coase theorem', *Journal of political Economy* **98**(6), 1325–1348.

Keasey, K. & Moon, P. (1996), 'Gambling with the house money in capital expenditure decisions', *Economics Letters* **50**, 105–110.

Loomes, G., Moffatt, P. G. & Sugden, R. (2002), 'A microeconometric test of alternative stochastic theories of risky choice', *Journal of Risk and Uncertainty* **24**, 103–130.

Maddala, G. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York.

Moffatt, P. G. & Peters, S. A. (2001), 'Testing for the presence of a tremble in economic experiments', *Experimental Economics* **4**, 221–228.

Nagel, R. (1995), 'Unravelling in guessing games: an experimental study', *American Economic Review* **85**, 1313–1326.

Nelson, F. D. (1976), 'On a general computer algorithm for the analysis of models with limited dependent variables', *Annals of Economic and Social Measurement* **5**, 493–509.

Plott, C. R. & Zeiler, K. (2007), 'Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory?', *American Economic Review* **97**(4), 1449–1466.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. (1991), 'Bargaining and market behaviour in jerusalem, ljubljana, pittsburgh and tokyo: An experimental study', *American Economic Review* **81**, 1068–1095.

Siegel, S. & Castellan, N. J. (1988), *Non-parametric statistics for the behavioral sciences, 2nd edition*, McGraw Hill.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, CRC Press.

Solnick, S. (2001), 'Gender differences in the ultimatum game', *Economic Inquiry* **39**, 189–200.

Thaler, R. & Johnson, E. (1990), 'Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice', *Management Science* **36**, 643–660.