Hypothesis testing using complex survey data

A Short Course presented by

Peter Lynn, University of Essex

in association with the conference of the European Survey Research Association

Prague, 25 June 2007

1. Objective: Simple Hypothesis Tests

Survey data are often used to test hypotheses. Hypotheses of interest are typically complex, involving several variables, for example:

- Differences in pay between men and women in urban areas can be explained by differences in occupation, hours worked and length of time in post

But in this course the examples we will use will be simple hypotheses. The ideas extend to more complex hypotheses.

Consider the following question, which is asked on the European Social Survey (ESS):

Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.

You can't be too careful										***Most people can be trusted***	**(Don't know)**
00	01	02	03	04	05	06	07	08	09	10	88

We might be interested in whether the mean score given in reply to this question (ppltrst) differs between nations. If the mean score is higher in one nation than another, then we might conclude that people in the first nation are more trusting than people in the second nation.

The mean scores given in the Czech Republic (CZ), Hungary (HU), Slovenia (SI), France (FR) and Portugal (PT) by ESS round 1 respondents (2002-03) were as follows:

```
CZ |   4.2543
FR |   4.4673
HU |   4.0794
PT |   4.0007
SI |   3.9768
```

It would appear that the French are the most trusting amongst these five nations, with the Slovenians the least trusting. But are these differences in means "significant"? In other words, are we confident that they reflect true differences in means between the respective populations as a whole?

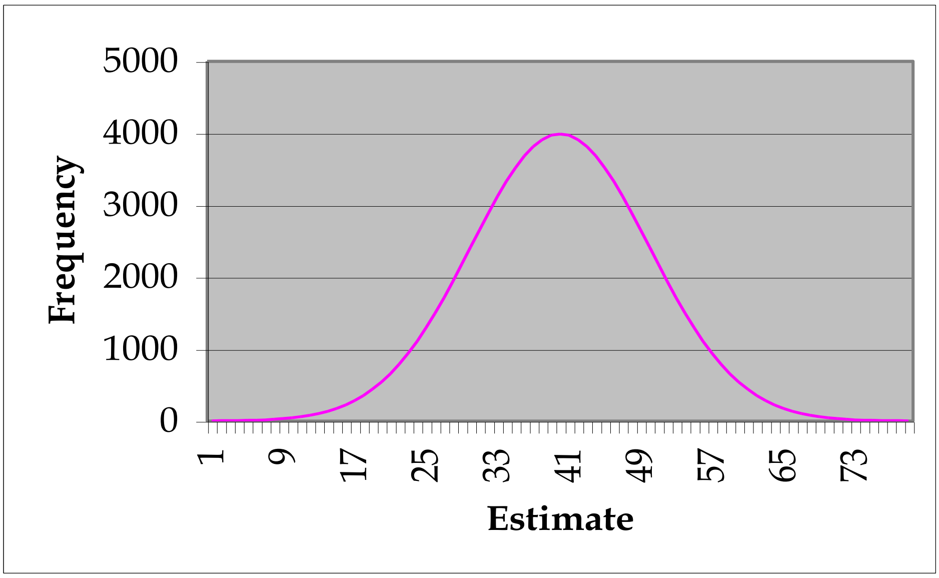
To answer this question, we need more information than just the means themselves. To see just what information we need, we must consider sampling theory.

2. Revision of some basic sampling theory

Sampling theory allows us to make statements about the “precision” of a sample estimate. Essentially, these are statements about how likely it is that a sample estimate falls within a particular distance of the true population value of which it is an estimate. This likelihood - or probability – depends solely on the sample design.

A sample design, D, defines a large set of possible samples that could be selected. For a particular estimator, E – e.g. mean score on the ESS trust question – each of those samples will provide an estimate. The estimates will vary over the samples.

The complete set of possible estimates is known as the “sampling distribution” of estimator E under sample design D. For most sample designs used in social surveys and for many of the kinds of estimators in which we are typically interested, sampling distributions are approximately “normally distributed”, meaning that they have a bell shape:



The normal distribution has some useful properties. It is symmetric. And there is a known relationship between the distance from the centre of the distribution (in terms of standard deviations) and the proportion of the area under the curve covered. For example, plus or minus 1.96 standard deviations covers 95% of the area under the curve.

In the case of a sampling distribution, this means that 95% of the samples that might be selected under design D will produce an estimate that is within 1.96 standard deviations of the true population value (assuming that the sampling distribution is centred on the true value). So, to make a precision statement of the form, “there is a 95% chance that the true value is within plus or minus *z* units of our sample estimate”, we need only to be able to estimate the standard deviation of the sampling distribution of the estimator – otherwise known as the standard error of the estimate. This is the extra information that we need in order to assess whether observed differences in means are significant.

Let’s consider the case of “simple random sampling” (SRS). It is a somewhat artificial case as SRS is rarely used in practice. But it is useful, for three reasons:

- The theory is relatively simple, so it is a comfortable place to start;
- SRS provides a standard design which we can use as a “benchmark”, against which to compare other – more realistic - designs;
- Much data analysis software carries out calculations under the assumption that the data are from a SRS – either by default or as the only option. We should try to understand what our software is doing.

SRS is a sample design where every unit in the study population has an equal, and independent, probability of selection. Note that many of the features often used in practical sample design, such as stratification, clustering and the use of variable sampling fractions, are not permissible within the definition of SRS. Stratification and clustering both cause selection probabilities to be dependent; variable sampling fractions cause selection probabilities to be unequal.

If we select a SRS of *n* units from a population of *N* units, the (sampling) variance of the sample mean of a variable *y* will be:

$$Var(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) \quad - \qquad (1)$$

where $S^2 = Var(y) = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

In most data analysis software, if you request the variance of a mean, this is the quantity that will be estimated (by default). In fact, the term $\left(1 - \frac{n}{N}\right)$ - known as the “finite population correction” - will almost certainly be ignored, as the software does not know *N*, the size of the population. Ignoring this term usually makes no difference as the value of this term is typically very close to 1.0. And S^2 will most likely be estimated by its sample analogue, s^2. So the estimate provided by the software will be:

$$V\hat{a}r(\bar{y}) = \frac{s^2}{n} \quad - \qquad (2)$$

The standard error is the square root of the variance, so the estimated standard error is simply the square root of the estimated variance as in (2).

3. Testing Differences in Mean Scores

The estimated standard errors of the mean trust scores (assuming SRS) are:

```
      Nation |      Mean   Std. Err.

          CZ |    4.2543   .06590
          FR |    4.4673   .05821
          HU |    4.0794   .05838
          PT |    4.0007   .05927
          SI |    3.9768   .06498
```

So now we can estimate 95% confidence intervals around the means, as these are plus or minus 1.96 standard errors. Our software gives us:

```
      Nation |      Mean   Std. Err.    [95% Conf. Interval]
-------------+------------------------------------------
          CZ |    4.2543   .06590       4.1251    4.3834
          FR |    4.4673   .05821       4.3532    4.5814
          HU |    4.0794   .05838       3.9649    4.1938
          PT |    4.0007   .05927       3.8845    4.1168
          SI |    3.9768   .06498       3.8495    4.1042
```

But how does this help us to assess whether the means are different from one another? Well, if we compare the confidence intervals for CZ and SI we see that they do not overlap at all. So it seems very unlikely that the true values for those two countries are the same. But if we compare, say, CZ and FR we find that the intervals overlap (slightly). So we still cannot be sure whether the difference is significant.

We need to state a formal hypothesis. We usually do this in terms of a “null hypothesis”, for example:

H_0: $\bar{Y}_{CZ} = \bar{Y}_{FR}$

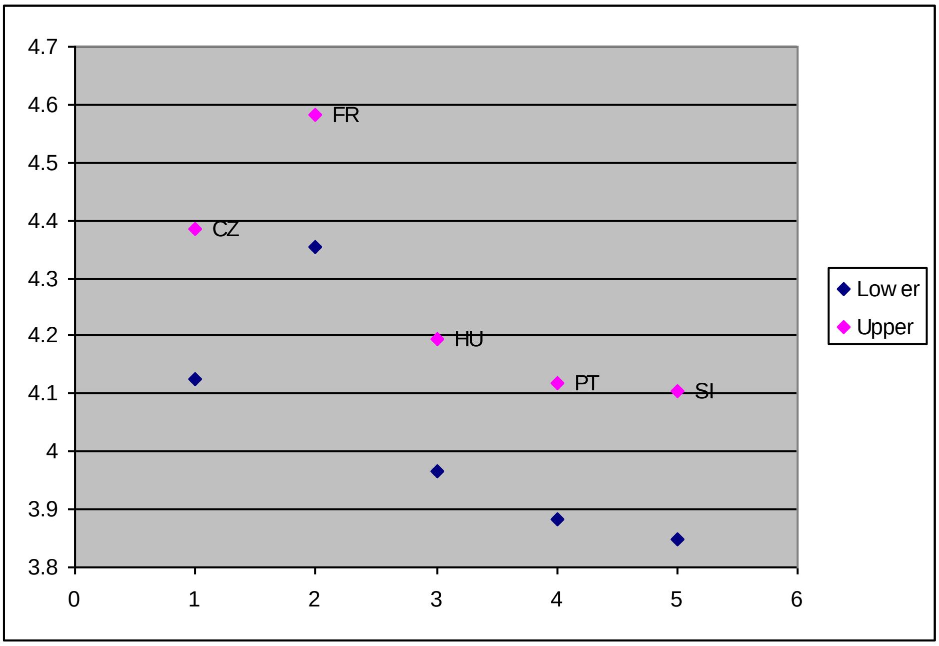
We then carry out a test to determine whether the data contain evidence to reject the null hypothesis. If the test rejects the null hypothesis then we would say that we have evidence that the means for CZ and FR differ.

An appropriate test for a difference in means is a “Wald test”. We can ask our software to perform this for us:

```
  [ppltrst]CZ - [ppltrst]FR = 0: F(1, 30970) = 5.87;  Prob > F = 0.0154
```

So, there appears to be a probability of only 0.0154, or 1.54%, that we would have observed a difference in means at least as large as the one actually observed, if the true means were the same. We might say that at the 0.05 level we would reject the null hypothesis of equal means in CZ and FR. So, the French are more trusting than the Czechs!

The figure below shows the estimated confidence intervals for the mean trust score for all five nations:



F-test results of comparisons between CZ and each of the other countries are as follows:

```
  [ppltrst]CZ - [ppltrst]FR = 0: F(1, 30970) = 5.87;  Prob > F = 0.0154
  [ppltrst]CZ - [ppltrst]HU = 0: F(1, 30970) = 3.95;  Prob > F = 0.0470
  [ppltrst]CZ - [ppltrst]PT = 0: F(1, 30970) = 8.19;  Prob > F = 0.0042
  [ppltrst]CZ - [ppltrst]SI = 0: F(1, 30970) = 8.99;  Prob > F = 0.0027
```

4. Variable Sampling Fractions

However, the estimates presented so far all assume that the sample in each nation is SRS. In fact, the ESS sample design is not SRS in any of these nations (see Lynn et al 2007).

HU and SI both selected their ESS round 1 sample from their national population register, enabling them to select persons with equal probabilities. But the other three nations used sample designs in which selection probabilities varied between units (persons). In all three cases the units listed and selected were addresses or households rather than persons. Then, in the field, interviewers would randomly select one person at the address to interview. This results in persons living alone having greater selection probabilities than persons living in 2-person households, etc.

When a sample design involves variable sampling fractions, “design weights” should be used in order to permit “design-unbiased” estimation. Design weights simply make each observation contribute to the estimate in inverse proportion to its selection probability. If households were selected with equal probabilities and then one person selected at random at each household then, compared to persons living alone, those living in 2-person households would receive a relative design weight of 2.0, those in 3-person households a weight of 3.0, and so on.

A weighted sample mean (estimate of population mean) would be calculated as follows:

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i}^{n} w_i} \quad - \qquad (3)$$

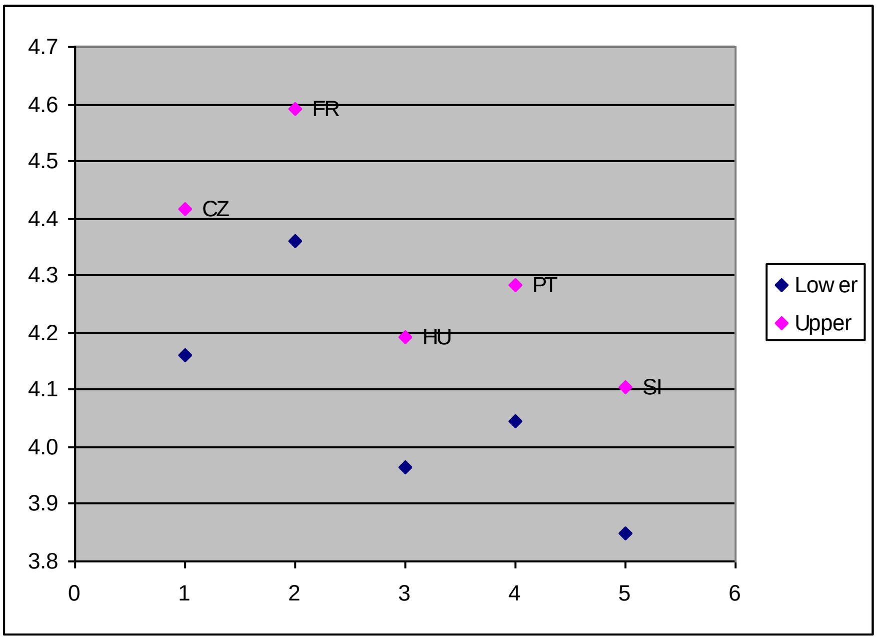
We should take the design weights into account in estimating the mean trust scores. If we ask our software to estimate means using (3), having specified which variable on the data set contains the design weight, w_i, we obtain:

```
      Nation |      Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
          CZ |    4.2889    .06519       4.1611     4.4167
          FR |    4.4759    .05811       4.3620     4.5898
          HU |    4.0794    .05838       3.9649     4.1938
          PT |    4.1638    .06033       4.0456     4.2821
          SI |    3.9768    .06498       3.8495     4.1042
```

Note that the estimates for both HU and SI are exactly the same as before, but for the other three nations both the estimate of the mean and the width of the confidence intervals have changed. These changes also affect the results of our tests of differences, which are now as follows:

```
  [ppltrst]CZ - [ppltrst]FR = 0: F(1, 30970) = 4.59;  Prob > F = 0.0322
  [ppltrst]CZ - [ppltrst]HU = 0: F(1, 30970) = 5.73;  Prob > F = 0.0167
  [ppltrst]CZ - [ppltrst]PT = 0: F(1, 30970) = 1.98;  Prob > F = 0.1593
  [ppltrst]CZ - [ppltrst]SI = 0: F(1, 30970) =11.49;  Prob > F = 0.0007
```

It seems that by ignoring the design weights, as we did earlier, we were over-estimating the significance of the differences between CZ and both FR and PT, but under-estimating the significance of the difference between CZ and HU. This can be seen in the plot of the estimated confidence intervals, using weighted data:



The intervals for both FR and PT now overlap with that for CZ more than before, while the interval for HU overlaps with CZ less than before.

5. Some More Sampling Theory

In fact, design weights affect not only estimates of means but also the variance of those estimates. This can be seen in the expression for the variance of a mean under stratified simple random sampling, as we can think of the weighting classes as strata (compare this with expression (1)):

$$Var(\bar{y})=\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{N^2 n_h}\left(1-\frac{n_h}{N_h}\right) \qquad - \qquad (4)$$

Note that the design weights are $w_h = \frac{N_h}{n_h}$ and that $N = \sum_{h=1}^{H} N_h$, so (if we ignore the finite population corrections), we can rewrite this as:

$$Var(\bar{y})=\frac{\sum_{h=1}^{H} w_h^2 n_h S_h^2}{\left(\sum_{h=1}^{H} w_h n_h\right)^2} \qquad - \qquad (5)$$

This can be estimated from the survey data provided we know the design weights for each sample unit (the s_h^2 provide estimates of S_h^2).

We can ask our software to estimate standard errors and confidence intervals taking into account the design weights:

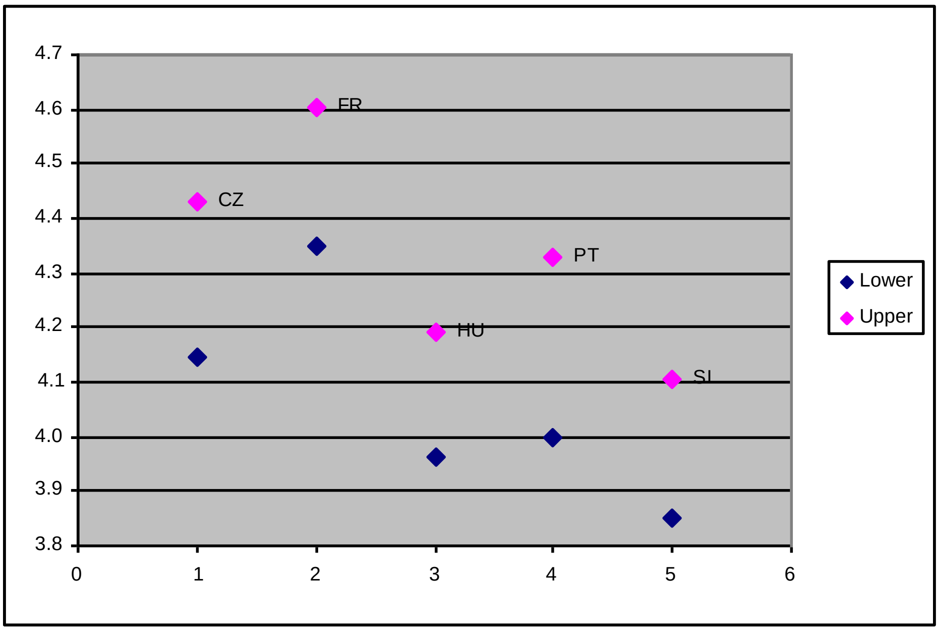
```
      Nation |      Mean   Std. Err.    [95% Conf. Interval]
-------------+-------------------------------------------------
          CZ |    4.2889   .07258       4.1466    4.4311
          FR |    4.4759   .06456       4.3494    4.6025
          HU |    4.0794   .05837       3.9650    4.1938
          PT |    4.1638   .08387       3.9995    4.3282
          SI |    3.9768   .06496       3.8495    4.1041
---------------------------------------------------------------
```

Note that the estimates of standard error are now larger than in the previous analysis for the three nations that do not have equal-probability designs. The standard error estimate has increased by a factor of 1.39 for PT, 1.11 for FR and 1.11 for CZ. These factors may be referred to as “mis-specification factors”: the factor by which the standard error is under-estimated due to mis-specifying the data structure. The mis-specification factor is closely related to, though not identical to, the “design factor”. The design factor due to the use of variable sampling fractions is the increase in standard errors relative to a SRS.

The tests of differences are now as follows:

```
 [ppltrst]CZ - [ppltrst]FR = 0: F(1, 30970) = 3.71;  Prob > F = 0.0542
 [ppltrst]CZ - [ppltrst]HU = 0: F(1, 30970) = 5.06;  Prob > F = 0.0245
 [ppltrst]CZ - [ppltrst]PT = 0: F(1, 30970) = 1.27;  Prob > F = 0.2597
 [ppltrst]CZ - [ppltrst]SI = 0: F(1, 30970) =10.26;  Prob > F = 0.0014
```

The P-values have increased in all cases. In particular, the P-value for the CZ-FR difference is now larger than 0.05, so we would no longer reject at this level the null hypothesis of equal means in CZ and FR. Remember that the P-value for this comparison was only 0.015 in our initial analysis where we ignored design weights completely. Again, we can see this graphically, as the confidence intervals for CZ and FR clearly overlap more than in the previous analyses:



6. Clustering

The use of variable sampling fractions (and hence design weights) is not the only way in which the ESS sample designs differ from SRS. In all five countries, multi-stage samples are selected, resulting in samples that are "clustered". This has the potential to affect standard errors of estimates. In general, if clusters are more homogeneous than the overall population, which is often the case, sample clustering will increase the size of standard errors.

The form of the variance of a mean gets complicated if we have both variable sampling fractions and a multi-stage clustered design (see, e.g., StataCorp 2005, p.261), but the approximate effect of a clustered design is to increase the variance by a factor of:

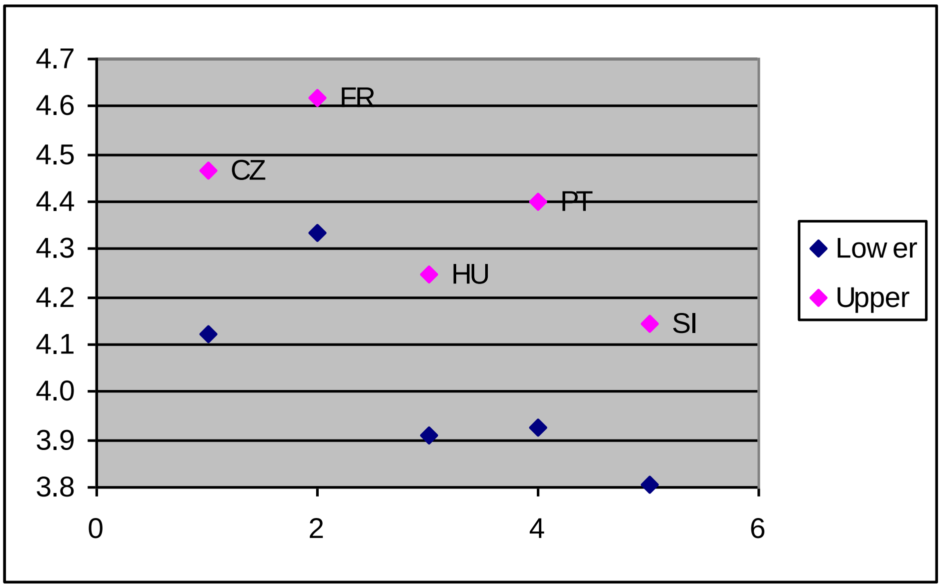
$$Deff_{cy} = 1 + \left(b^* - 1\right)\rho_y \qquad - \qquad (6)$$

where b^* is a weighted mean cluster sample size and ρ_y is the intra-cluster correlation for y (see Kish 1965, pp.170-171; Lynn & Gabler 2005).

If we ask our software to take into account the sample clustering as well as the design weights, we get the following estimates:

```
      Nation |      Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
          CZ |    4.2889   .08774       4.1169    4.4611
          FR |    4.4759   .07356       4.3319    4.6200
          HU |    4.0794   .08558       3.9116    4.2471
          PT |    4.1638   .12063       3.9273    4.4003
          SI |    3.9768   .08622       3.8078    4.1459
--------------------------------------------------------------
```

```
  [ppltrst]CZ - [ppltrst]FR = 0: F(1, 30970) = 2.65;  Prob > F = 0.1036
  [ppltrst]CZ - [ppltrst]HU = 0: F(1, 30970) = 3.04;  Prob > F = 0.0812
  [ppltrst]CZ - [ppltrst]PT = 0: F(1, 30970) = 0.73;  Prob > F = 0.3933
  [ppltrst]CZ - [ppltrst]SI = 0: F(1, 30970) = 6.55;  Prob > F = 0.0155
```



What we observe is that if we take the relevant features of the sample design into account, the mean for CZ is not significantly different from the mean for FR, HU or PT at the 0.05 level. It is different from the mean for SI at the 0.05 level, but not at the 0.01 level.

This contrasts sharply with the results that we obtained with our naïve analysis, assuming SRS. In that case it seemed that all four of the differences were significant at the 0.05 level and two of them at the 0.01 level.

Taking the sample design correctly into account alters the conclusions! Furthermore, we have seen that the differences in the estimates of standard errors are partly due to the effect of variable sampling fractions and partly due to the effect of clustering of the sampling – so it is important to take *both* these factors into account.

7. Another Example

Another question on the ESS (pplhlp) has a similar structure to the one analysed above, but a different topic:

Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?

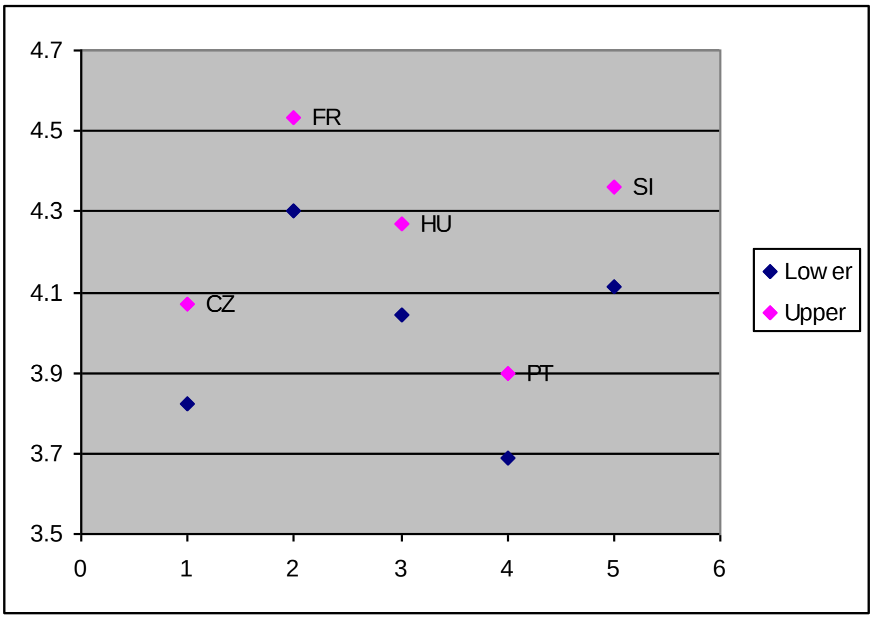
Most people would try to take advantage of me										***Most people would try to be fair***	***(Don't know)***
00	01	02	03	04	05	06	07	08	09	10	88

If we run equivalent analyses to those presented above, again using ESS round 1 data, we obtain the following results:

7.1: Results assuming SRS

```
      Nation |      Mean   Std. Err.    [95% Conf. Interval]
-------------+--------------------------------------------------
          CZ |    3.9471   .06319       3.8233    4.0710
          FR |    4.4175   .05973       4.3004    4.5346
          HU |    4.1556   .05754       4.0429    4.2684
          PT |    3.7926   .05444       3.6859    3.8993
          SI |    4.2389   .06359       4.1143    4.3635
----------------------------------------------------------------

  [pplhlp]CZ - [pplhlp]FR = 0: F(1, 30970) = 29.26;  Prob > F = 0.0000
  [pplhlp]CZ - [pplhlp]HU = 0: F(1, 30970) =  5.95;  Prob > F = 0.0147
  [pplhlp]CZ - [pplhlp]PT = 0: F(1, 30970) =  3.63;  Prob > F = 0.0640
  [pplhlp]CZ - [pplhlp]SI = 0: F(1, 30970) = 10.59;  Prob > F = 0.0011
```

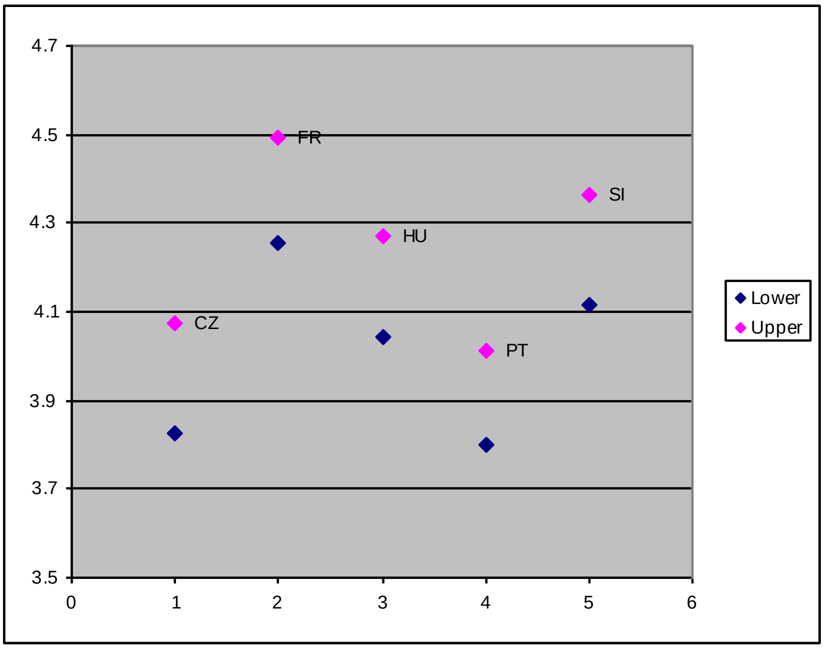


Differences between the mean for CZ and FR and SI appear highly significant (P<0.01); the difference with HU appears significant at the 0.05 level (P=0.015) and the difference with PT is almost significant at the 0.05 level (P=0.064).

7.2: Results using weighted means but assuming SRS in variance estimation

```
      Nation |      Mean   Std. Err.    [95% Conf. Interval]
-------------+----------------------------------------------
          CZ |    3.9491   .06307       3.8255     4.0728
          FR |    4.3740   .06013       4.2561     4.4919
          HU |    4.1556   .05754       4.0429     4.2684
          PT |    3.9059   .05507       3.7980     4.0139
          SI |    4.2389   .06359       4.1143     4.3635
---------------------------------------------------------------

  [pplhlp]CZ - [pplhlp]FR = 0: F(1, 30970) = 23.77;  Prob > F = 0.0000
  [pplhlp]CZ - [pplhlp]HU = 0: F(1, 30970) =  5.85;  Prob > F = 0.0156
  [pplhlp]CZ - [pplhlp]PT = 0: F(1, 30970) =  0.27;  Prob > F = 0.6058
  [pplhlp]CZ - [pplhlp]SI = 0: F(1, 30970) = 10.47;  Prob > F = 0.0012
```

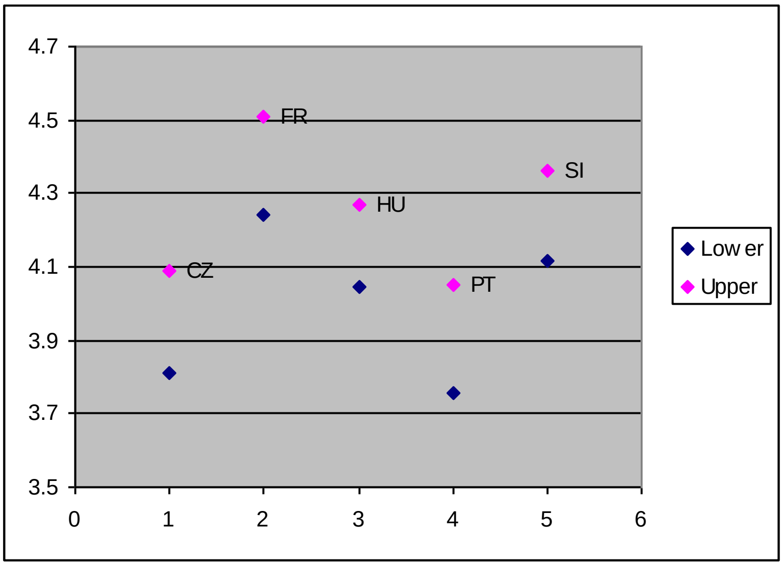


The main change here is that the weighted mean for PT is higher than the unweighted mean, with the result that the mean for PT no longer appears significantly different from that for CZ.

7.3: Results taking account of weighting, but not clustering, in variance estimation

```
      Nation |      Mean   Std. Err.    [95% Conf. Interval]
-------------+--------------------------------------------
          CZ |    3.9491   .07049       3.8110     4.0873
          FR |    4.3740   .06777       4.2412     4.5068
          HU |    4.1556   .05753       4.0429     4.2684
          PT |    3.9059   .07521       3.7585     4.0533
          SI |    4.2389   .06357       4.1143     4.3635
-------------------------------------------------------------
```

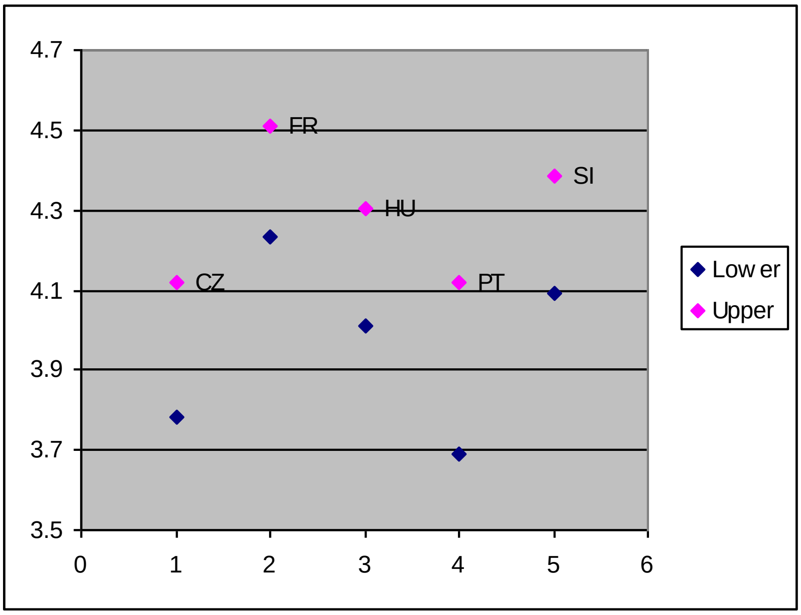
```
 [pplhlp]CZ - [pplhlp]FR = 0: F(1, 30970) = 18.88;  Prob > F = 0.0000
 [pplhlp]CZ - [pplhlp]HU = 0: F(1, 30970) =  5.15;  Prob > F = 0.0232
 [pplhlp]CZ - [pplhlp]PT = 0: F(1, 30970) =  0.18;  Prob > F = 0.6750
 [pplhlp]CZ - [pplhlp]SI = 0: F(1, 30970) =  9.32;  Prob > F = 0.0023
```



P-values have increased for all four tests, but the differences are unlikely to affect conclusions.

7.4: Results taking account of both weighting and clustering

```
      Nation |      Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
          CZ |    3.9491   .08587       3.7808     4.1174
          FR |    4.3740   .07069       4.2354     4.5126
          HU |    4.1556   .07490       4.0088     4.3025
          PT |    3.9059   .10910       3.6920     4.1198
          SI |    4.2389   .07537       4.0911     4.3867
--------------------------------------------------------------
```



```
  [pplhlp]CZ - [pplhlp]FR = 0: F(1, 30970) = 15.38;  Prob > F = 0.0001
  [pplhlp]CZ - [pplhlp]HU = 0: F(1, 30970) =  3.20;  Prob > F = 0.0739
  [pplhlp]CZ - [pplhlp]PT = 0: F(1, 30970) =  0.10;  Prob > F = 0.7523
  [pplhlp]CZ - [pplhlp]SI = 0: F(1, 30970) =  6.39;  Prob > F = 0.0115
```

In this example, the most dramatic impact of mis-specification is to over-state the difference in means between CZ and PT. However, this is mainly caused by failure to apply design weights in estimating the mean: the analysis in section 7.2 already showed no significant difference between CZ and PT, even without taking the design into account.

The other noticeable impact of mis-specification is to over-state the evidence of a difference between CZ and HU. This is caused entirely by the failure to estimate the variance of the estimates correctly (P=0.016 in 7.2, cf. P=0.074 in 7.4).

8. A Third Example: Change Between Rounds

Here we are interested in testing whether the mean score changes between rounds 1 (2002-03) and 2 (2004-05) of the ESS. We carry out the estimation in the same four ways as previously, for the variable ppltrst for Luxembourg:

8.1: Results assuming SRS

```
       Round |      Mean   Std. Err.   [95% Conf. Interval]
-------------+--------------------------------------------------
           1 |    5.2133   .05871      5.0982    5.3284
           2 |    5.0080   .06093      4.8885    4.1274
----------------------------------------------------------------

  [ppltrst]1 - [ppltrst]2 = 0: F(1, 30970) =  5.89;  Prob > F = 0.0153
```

8.2: Results using weighted means but assuming SRS in variance estimation

```
       Round |      Mean   Std. Err.   [95% Conf. Interval]
-------------+--------------------------------------------------
           1 |    5.1848   .05846      5.0701    5.2994
           2 |    5.0152   .06065      4.8963    5.1342
----------------------------------------------------------------

  [ppltrst]1 - [ppltrst]2 = 0: F(1, 30970) =  4.05;  Prob > F = 0.0443
```

8.3: Results taking account of weighting, but not clustering, in variance estimation

```
       Round |      Mean   Std. Err.   [95% Conf. Interval]
-------------+--------------------------------------------------
           1 |    5.1848   .06519      5.0570    5.3126
           2 |    5.0152   .07456      4.8691    5.1614
----------------------------------------------------------------

  [ppltrst]1 - [ppltrst]2 = 0: F(1, 30970) =  2.93;  Prob > F = 0.0870
```

The sample design in Luxembourg was unclustered, so there is no need to take into account clustering. In this example, the test of a difference in means, correctly taking into account the sample design, provides no evidence at the 0.05 level of a difference (P=0.087). But ignoring the weights in variance estimation would suggest evidence of a reduction in trusting between ESS rounds 1 and 2 (P=0.044). And additionally ignoring the weights in estimating the means would suggest even stronger evidence of a reduction (P=0.015).

9. Some Comments on Software Implementation

The analyses presented here were carried out in Stata. The commands are quite simple to implement, using the SVY commands to take into account the sample design. It is necessary to have a variable that contains the design weight (`dweight`) and a variable that indicates the cluster, or "primary sampling unit" (`psunit`).

For comparing the mean of ppltrst between the five countries:

```
svyset [pw = dweight], psu(psunit)
svy: mean ppltrst if (set==1 & essround==1), over(cntcode)
test [ppltrst]4 = [ppltrst]9
test [ppltrst]4 = [ppltrst]12
test [ppltrst]4 = [ppltrst]19
test [ppltrst]4 = [ppltrst]21
```

For comparing the mean of ppltrst between rounds 1 and 2 for Luxembourg:

```
svyset [pw = dweight]
svy: mean ppltrst  if cntcode==15, over(essround)
test [ppltrst]1 = [ppltrst]2
```

Similar commands are available in SPSS (in the 'Advanced Statistics' module) and in SUDAAN.

References

Kish L (1965) *Survey Sampling*. New York: Wiley.

Lynn P & Gabler S (2005) Approximations to b^* in the prediction of design effects due to clustering, *Survey Methodology* 31, 101-104.

Lynn P, Häder S, Gabler S & Laaksonen S (2007) Methods for achieving equivalence of samples in cross-national surveys: the European Social Survey experience, *Journal of Official Statistics* 23, 107-124.

StataCorp (2005) *Stata Survey Data Reference Manual Release 9*. Stata Press: College Station, Texas.