Annex 1. Further information on the statistical approach employed in the Fourth Periodic Evaluation

Introduction

This section summarises the key points relating to the statistical treatment of radioactive discharges data as set out in the Fourth Periodic Evaluation. Previous Periodic Evaluation Reports already devoted large sections to this issue. In particular, Annex 1 of the Third Periodic Evaluation report (OSPAR, 2009a) described in detail the statistical methods applicable to the OSPAR radioactive substances strategy, in agreement with the recommendations from ICG-Stats (OSPAR, 2009b). Therefore, this information is not repeated here but the tools used in the present report, devoted to radioactive discharges only, are recalled.

The baseline elements

RSC has defined the baseline period for discharges from the nuclear sector as the period 1995 to 2001 and the baseline period for discharges from the non-nuclear oil and gas sub-sector as 2005 to 2011. The baseline value is the mean (average) of the summed discharges for each of the years of the relevant baseline period for a particular indicator (individual or grouping of radionuclides) for a particular Contracting Party or sub-sector. A predicted interval (PI) centered on this mean of 1.96 times the standard deviation, gives the baseline bracket, where the “bracket” would contain 95% of the observed discharge values, provided they were normally distributed. Comparing subsequent data with a baseline (especially a baseline constructed from several years’ data) is fundamentally different from examining data for the presence and nature of a trend. The nearest statistical analogy is with considering samples to see whether they have been drawn from statistically significantly different populations. In effect, the data from which the baseline has been derived are considered as one sample, and the question is whether the subsequent data are from the same population or from a different one. If the true levels of discharges have changed sufficiently, they can be regarded as a different population. In other words, defining the baseline period as a sample of 7 annual values (1995 to 2001 or 2005 to 2011) described by its mean (baseline value) and its 95% PI (baseline bracket) is a way to embed the changes in the annual values as fluctuations over the baseline period rather than a true temporal trend.

The assessment elements

To assess the changes between new discharges and the baseline period, a second sample of observations must be considered i.e. the assessment period. In this Fourth Periodic Evaluation report, the assessment period for discharges from the nuclear sector is the period 2007 to 2013. The assessment value is the mean (average) of the summed discharges for each of the years of the assessment period for a particular indicator (individual or grouping of radionuclides) for a particular Contracting Party or sub-sector.

Comparison of the baseline and assessment elements

To compare the baseline period and the assessment period, it is assumed that these two samples are made of independent observations. A two-step methodology is then applied:

i) Simple comparison

The assessment value for the assessment period is compared against the baseline value and baseline bracket. The idea is that if there has been no statistically significant change in the levels of discharge, there is a good probability that subsequent observations (the assessment value) will fall within PI, whereas a true change in these levels should result in subsequent observations (the assessment value) lying outside of this interval. This simple comparison method is not very sensitive and it includes a major risk of type I error (a difference is detected when there is no statistically significant difference). This method should therefore only be used as a first simple indicator for the comparison of the assessment period with the baseline period. This simple comparison cannot produce “statistically significant” results.

Furthermore, the lower baseline bracket calculation could yield a negative number (in such cases, the lower baseline bracket is set to zero) which would impair this approach. However, since it was included in the baseline elements as agreed by the 2003 Ministerial Meeting of the OSPAR Commission, it has been retained as a method of comparison.

ii) Statistical tests

An efficient comparison of two samples to detect a potential difference between them requires us to treat the second sample as a whole, and to use all the information it contains.

Comparison methods can be divided into parametric and non-parametric methods. The difference between the two types of methods is that parametric methods need to make assumptions about the nature (parameters) of the two data sets that are being compared, whereas non-parametric methods do not. The assumption that is most usually made in parametric methods of comparison is that both samples are drawn from two populations where the variables are independent and share an identical normal distribution.

The classic parametric test for whether two samples are drawn from populations with different characteristics is one of the forms of the Student's t test. The comparison is between the observations in the baseline period (1995 to 2001), and the observations of the assessment period (2007 to 2013). The comparison being made here is between two populations where the members of each population are not related to each other and therefore it is the unpaired test which is appropriate. There are two forms of the unpaired test: the homoscedastic, where the variances of the two populations are (or are assumed to be) the same, and the heteroscedastic, where the variances of the two populations are (or are assumed to be) different. Using the homoscedastic form if the variances are not the same could lead to a Type I error. In the comparison being made here, there is no reason to think that the variances are the same; therefore, the heteroscedastic form (the Student's t Welch Aspin test) was more appropriate, and has been chosen for use as the parametric test. However, this test in its general form may not be entirely suited to the present evaluation, since there can be no discharge values less than zero (the distribution of the data may be a truncated normal distribution) and the number of data points available in the sample is very small.

For the non-parametric comparison methods, a widely used method is the Mann-Whitney test. This test belongs to the wider family of the rank tests which comprise Kendall's Tau or S test. This group of tests is most appropriate when it is desired to see whether the means of two samples represent different populations and no assumption is (or can be) made on how the observations are distributed. The Mann-Whitney test is carried out by ranking the combined data set of the two samples in ascending sequence, and assigning a rank to each data element (1, 2, 3...), irrespective of the sample to which it belongs. If two or more data elements are equal, they are given their average rank. The ranks for the data elements of the smaller of the two samples (or either sample if they have the same number of data elements) are then summed to give the rank-sum. For small samples, the rank sum is compared to what would result if the data were ranked in a single data set and assigned at random to two groups having the same number of observations as the original samples. The random-assignment calculation gives a probability o for any given rank sum for two samples of the given sizes. If the probability o for the rank sum calculated is less than the chosen probability cut-off level (normally 0.05, or 5%), then the null hypothesis should be rejected, and the conclusion should be that the two samples are from different populations. Ranking tests are more robust than parametric methods: that is to say, they are more likely to lead to Type II errors than to Type I errors that should be prevented.

Annual values for the assessment period are compared against annual values for the baseline period using both the Student's t Welch-Aspin test and the Mann-Whitney test to qualify conclusions drawn from simple comparisons between the assessment value and the baseline value and baseline bracket. Where P values of less than 0.050 have been determined using these statistical tests, the difference between annual values for the baseline period and assessment period can be said to be ‘statistically significant’.

The outcome of the simple comparison and the statistical tests allows for the following conclusions to be made:

- Where both statistical tests give results that are statistically significant it can be concluded that there is evidence of a reduction or increase (as indicated by the simple comparison) in the discharge between the assessment and baseline periods.
- Where only one statistical test gives a result that is statistically significant it can be concluded that there is some evidence of a reduction or increase (as indicated by the simple comparison) in the discharge between the assessment and baseline periods.
- Where neither statistical test gives a result that is statistically significant it can be concluded that there is no evidence of any change in the discharge between the assessment and baseline periods.

Tutorial

All statistical analysis were performed using the R Environment Software for Windows. This software can be downloaded and used free of charge from the R-Cran web site @ http://www.r-project.org/.

The following colours are used in this tutorial for clarity

```
Comments appear in green, they start with hashtag character #
R commands appear in red, they can be copied and pasted at the R
console prompt >
R outputs appear in blue
```

- Start R
- Input the discharge data (here French total beta (excluding tritium) from the nuclear power sub-sector is given as an example)

Year	Discharges (TBq)
1995	0.0921
1996	0.068
1997	0.0546
1998	0.0415
1999	0.0381
2000	0.03
2001	0.0305
2007	0.00951
2008	0.008728
2009	0.007686
2010	0.008372
2011	0.007302
2012	0.009839
2013	0.012272

- Execute the following R commands

```
# enter the series of annual discharges for the baseline [1995-2001].
> AnnualData.baseline<-c(0.0921, 0.068, 0.0546, 0.0415, 0.0381, 0.03,
0.0305)
# calculate the mean as the baseline.
> mean(AnnualData.baseline)
[1] 0.05068571
# calculate the upper bracket of the baseline.
> mean(AnnualData.baseline)-1.96*sd(AnnualData.baseline)
[1] 0.006074084
# calculate the lower bracket the baseline.
> mean(AnnualData.baseline)+1.96*sd(AnnualData.baseline)
[1] 0.09529734
# enter the series of annual discharges for the assessment period
[2007-2013].
> AnnualData.assessment<-c(0.00951, 0.008728, 0.007686, 0.008372,
0.007302, 0.009839, 0.0122718)
# calculate the mean as for the assessment period.
> mean(AnnualData.assessment)
[1] 0.009101257
# Compares the means of the baseline and the assessment period using a
Student t-test.
> t.test(AnnualData.baseline, AnnualData.assessment, var.equal=F)
# 'var.equal=F' parameter selects the heteroscedastic form of the
Student t.test: the Welch Aspin test.
        Welch Two Sample t-test

data:  AnnualData.baseline and AnnualData.assessment
t = 4.8209, df = 6.0644, p-value = 0.002853
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02053181 0.06263710
sample estimates:
  mean of x   mean of y
0.050685714 0.009101257
# Compares the baseline and the assessment period using a Rank test.
> library(exactRankTests)
> wilcox.exact(AnnualData.baseline, AnnualData.assessment)
        Exact Wilcoxon rank sum test

data:  AnnualData.baseline and AnnualData.assessment
W = 49, p-value = 0.0005828
alternative hypothesis: true mu is not equal to 0
```

France nuclear power sub-sector	Baseline value (TBq)	Lower baseline bracket (TBq)	Upper baseline bracket (TBq)	Assessment value (TBq)	Student's t Welch Aspin test (P value)	Mann-Whitney test (P value)
Total beta	5.07E-02	6.07E-03	9.53E-02	9.10E-03	2.85E-03	5.83E-04

(excluding tritium)						

The results are presented in the above table

In the above example, the assessment value is lower than the baseline value but not the lower baseline bracket. The P values from both statistical tests are less than 0.050, meaning that there is a statistically significant difference between the assessment period and the baseline period. Taken together, it can be concluded that there is evidence of a reduction in discharges between the baseline period and the assessment period.

References

OSPAR, 2009a. Third Periodic Evaluation of progress towards the objective of the Radioactive Substances Strategy. Publication 455/2009

OSPAR, 2009b. Assessment on Statistical techniques applicable to the OSPAR Radioactive Substances Strategy. Publication 454/2009

R Development Core Team. (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.