# ASSESSMENT 9

## QUESTION 1

##### Statistical modelling

This practical session assignment uses data available from [data.gov.uk](https://data.gov.uk/) about the results of the [National Paediatric Diabetes Audit](https://data.gov.uk/dataset/b532ab89-360d-47aa-99ea-d21eb01d3333/national-paediatric-diabetes-audit-2017-18) 2011-2017.

The original data were provided as a spreadsheets for each year of audit. From each, the tab pertaining to HbA1c or [glycated haemoglobin](https://www.diabetes.co.uk/what-is-hba1c.html) was extracted into a single spreadsheet (*npda.xlsx*), which is provided in the folder containing this practical session assignment. The variables are:

1. *unit* - paediatric diabetes unit (PDU) unique identifier
2. *region* - geographic region in which the PDU is located
3. *target1-target7* - proportion of cases where the HbA1c is less than the older target of 58mmol/mol (see p47 in [the Royal College of Paediatrics report](https://www.rcpch.ac.uk/sites/default/files/2019-05/NPDA-national-report-2017-18_v2-updated-2019-05-30_0.pdf))
4. *patients1-patients7* - number of cases of paediatric diabetes managed by the PDU

*Note*: *target1* and *patients1* refer to data reported in 2011/12. *Target2-7* and *patients2-7* refer to data reported during each successive year up to 2017/18.

###### Self assessment question

1. Import the data in the spreadsheet.
2. Check that each instance of the variable *unit* is unique.
3. Validate each of the remaining variables in a suitable manner appropriate to their type.

QUESTION 2

##### Exploratory data analysis

###### Results of validation

When the validation was originally carried out, two problems were identified. The first was that the unit number was not always of the form: PZ### where # is a numeric digit. The second was that the percentage of people with HbA1c less than 58 mmol/mol was recorded in different ways, resulting in different ranges for the proportion. The provided data set suffers from neither of these problems, having been generated after these problems were addressed.

###### Regional variation

For this practical session, the first investigation will be into whether there is a north/south divide, based on the region of the PDU, in the proportion of patients below the target of 58 mmol/mol, for each year.

###### Self assessment question

1. Create an informat which allocates region to three numeric categories, as follows:
   * London and South East = 1
   * South Central = 1
   * South West = 1
   * East Midlands = 2
   * East of England = 2
   * Wales = 2
   * West Midlands = 2
   * Yorkshire and Humber = 3
   * North East = 3
   * North West = 3
2. Create a format for the three categories, which prints:
   * 1 = north
   * 2 = middle
   * 3 = south
3. Create a new data set, which adds to the imported data, a new variable 'V', which converts the region into a numeric value using the informat, is labelled 'North/South' and uses the above format when it is output.   
   *Note*: 'V' should be given a suitable name.
4. For each year, tabulate:
   * number of PDUs
   * number with no proportion (missing data)
   * minimum
   * 25th percentile
   * median
   * 75th percentile
   * maximum
   * mean
   * standard deviation

against the value of 'V', ensuring the use of an appropriate format for percentages.

1. Review the table. Between which groups (north, middle, south) are there differences? Which year has the largest differences?

QUESTION 3

##### Statistical modelling - two categories of independant variable

Since the regions labelled 'middle' are neither completely in the north or the south, they will be excluded from the analysis. This leaves two groups (north and south) for which values of the proportion of patients with HbA1c less than 58 mmol/mol can be determined.

###### Self assessment question

Select an appropriate statistical model to test whether there is any difference, in any year, between the two groups (north and south) in regard of the proportion of patients with HbA1c less than 58 mmol/mol.

What are the assumptions of the selected model? Are they met? Does the extent to which they are met, invalidate any inference you make?

*Hint*: since there are two independent samples of numeric data, a possible test is the Student's t-test.

|  |  |
| --- | --- |
| **YEAR** | **INFERENCE** |
| 1 | The Pr>F from the Equality of Variance is 0.6516 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.0457 0.0112)  For hypothesis of comparing control and treatment, t-value =-1.20 and the p-value = 0.2311  Since the p-values are greater than 0.05 we accept the alternative hypothesis that the value of north is equal to south. |
| 2 | The Pr>F from the Equality of Variance is 0.0020 which is less than 0.05. Hence, we read the Satterthwaite Method of the result.  The confidence interval for control is (-0.0232 0.0337)  For hypothesis of comparing control and treatment, t-value =0.37 and the p-value = 0.7145  Since the p-values are less than 0.05 reject the null hypothesis that the north and south are equal. |
| 3 | The Pr>F from the Equality of Variance is 0.4205 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.00690 0.0435)  For hypothesis of comparing control and treatment, t-value =1.44 and the p-value = 0.1532  Since the p-values are greater than 0.05 so we accept the alternative hypothesis that the north and south are equal. |
| 4 | The Pr>F from the Equality of Variance is 0.8844 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.0210 0.0377)  For hypothesis of comparing control and treatment, t-value =0.56 and the p-value = 0.5734  Since the p-values are greater than 0.05 so we accept the alternative hypothesis that the north and south are equal. |
| 5 | The Pr>F from the Equality of Variance is 0.1551 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.0109 0.0485)  For hypothesis of comparing control and treatment, t-value =1.25 and the p-value = 0.2130  Since the p-values are greater than 0.05 so we accept the alternative hypothesis that the north and south are equal. |
| 6 | The Pr>F from the Equality of Variance is 0.4779 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.00231 0.0604)  For hypothesis of comparing control and treatment, t-value =2.14 and the p-value = 0.0346  Since the p-values are greater than 0.05 so we accept the alternative hypothesis that the north and south are equal. |
| 7 | The Pr>F from the Equality of Variance is 0.3740 which is greater than 0.05. Hence, we read the Pooled Method of the result.  The confidence interval for control is (-0.0189 0.0476)  For hypothesis of comparing control and treatment, t-value =0.85 and the p-value = 0.3946  Since the p-values are greater than 0.05 so we accept the alternative hypothesis that the north and south are equal. |

Sas.com. 2020. [online] Available at: <https://www.sas.com/content/dam/SAS/en\_ca/User%20Group%20Presentations/Calgary-User-Group/Yankovsky-ExploringProcTtest-Apr2015.pdf> [Accessed 3 April 2020].