Annex 1 – GSG Statistical Tools and Techniques

Purpose

The purpose of this document is to provide examples of some of the statistical tools and techniques used by Statisticians and Statistical Data Scientists. This is not an exhaustive list, but serves to demonstrate the areas in which we might expect the profession to be building its technical capability.

This list should not be used as a checklist for recruitment.

Introduction

As a professional analytical group within the civil service it is important that we commit to maintaining and building on our technical skills. The statistical tools and techniques listed within this document are key for the statistical profession, and demonstrate how Data Science is fundamental to the work that we do. Whilst this document has been developed primarily with the statistical profession in mind, it is acknowledged that other analytical professions may be able to align or draw benefits from it, as well as those who work with statistics but who are not aligned to a profession.

The statistical tools and techniques within this document are presented against the same 'statistical strands' that are contained within the GSG Competency Framework:

1. Acquiring data/Understanding customer needs;
2. Data analysis; and
3. Presenting and disseminating data effectively.

It also draws on the elements contained within the Generic Statistical Business Process Model (GSBPM), which lists the processes followed to produce a statistical output. The GSBPM was formulated jointly by UN/OECD/Eurostat and has been adopted by many worldwide National Statistical Institutes as a process model for the production of statistical outputs.

The statistical tools and techniques contained should not be considered as a checklist of statistical techniques that any individual member of the statistical profession is expected to know. The analytical professions as a whole need members who collectively have a good understanding of these techniques. For each technique we have given examples of their use across the GSS, whether this is in an official statistics release, a methodology paper or annex, or a presentation to a conference.

The list of techniques is not exhaustive. Statistics is an enormous discipline which has an impact on so many aspects of modern life, and there are many more techniques and skills that statisticians could use to draw insight from a dataset. For example, Bayesian statistical techniques and thinking do not feature explicitly in this list, but are likely to become more and more useful as we deal with larger and less structured data in the future. As a profession we should value all of these. But the techniques described here constitute a good starting point for a new colleague trying to decide what to learn more about, and a good target for the statistical profession (or analytical professions as a whole) to aim at collectively gaining a sound understanding of.

1. Acquiring data/Understanding customer needs

This statistical strand captures a wide range of statistical principles, tools and techniques that are generally required at the start of the statistical process and which require a breadth of knowledge in order to make the appropriate decisions about data acquisition early on in the statistical process. It will depend on each job role as to the mixture of areas required from this statistical strand, however a flavour is provided here.

This statistical strand maps onto the **Design, Build and Collect Phases of the GSBPM**. The following is covered:

- Identifying the needs for the statistics
- Confirming the needs of stakeholders
- Establishing the high level objectives of the statistical outputs
- Identifying the relevant concepts and variables for which data are required
- Checking the extent to which current data sources can meet these demands (this involves researching the data available including administrative and open data sources)
- Making full use of open data, extracting from the internet/open sources where possible.
- Preparing the business case to get approval to produce the statistics
- Designing the output (in accordance with customer needs)
- Designing variable descriptions
- Designing and building the mode of data collection
- Designing and creating the sampling frame (where required)
- Designing the processing and analysis methodology
- Building or enhancing dissemination components
- Configuring workflows
- Testing the processes
- Collecting the data and loading it into a suitable electronic environment for next stage.

Statistical Knowledge required

Acquiring Data • Pros and cons of using surveys, censuses, administrative data, open data) • Open data standards • Storage for administrative data and survey microdata • Sharing data (e.g. in the Virtual Microdata Lab, other data archives, data labs, Administrative Data Research Centres) • Awareness of legal issues around collecting and sharing data • Design of data collection mechanisms (questionnaires, technology for data acquisition, longitudinal and cross-sectional surveys)	• Mode effects (quality and cost trade-offs, mixed mode surveys) • Data matching (exact matching, probability matching, statistical matching, data linking) • Data quality control and monitoring (improving data quality, prevent data contamination) • Sampling design (simple random sampling, stratified sampling, probability proportional to size, cluster sampling, the Neyman allocation)
Open Data • NoSQL databases • Hadoop • Git and Git hub	• Machine learning • Sentiment analysis • Web scraping

Examples

Applications across the GSS	*Useful resources*
DfT: Presentation of administrative data, Traffic counts **DfT:** Technical report, Processing of National Travel Survey GPS Pilot Data **DWP:** Collecting and sharing data, Local authority data sharing guide **DWP:** Family Resources Survey 2010/11 (methodology chapter) **MoJ**-**DWP:** Experimental statistics from data share, Linking data on offenders with benefit, employment and income data **ONS:** GSS Methodology Series, Sample Design Options for an Integrated Household Survey **ONS:** Sampling a Matching Project to Establish the Linking Quality (GSS Survey Methodology Bulletin, no.72) **Scottish Government:** Scottish Population Surveys Centralised Weighting Project	• Open Government Project: data.gov.uk • Open Data Institute: Guides • Sampling Techniques; Cochrane, W.G. • Survey Sampling; Kish, L. • GSS Survey Methodology Bulletin (archived website) • Price and Quantity Index Numbers; Balk, B. M. • International Labour Organisation, Consumer Price Index Manual: Theory and Practice 2004 **Open Data** • GSS Data Blog

2. Data Analysis

In this statistical strand, the statistical outputs are produced, examined in detail and made ready for dissemination. The same principles apply regardless of how the data were sourced. It includes ensuring that the data analysis is 'fit for purpose' prior to dissemination to customers. It will depend on the job role as to the mix of knowledge required from this strand.

This statistical strand maps onto the **Process and Analyse phases of the GSBPM; aspects of quality are also considered**. The following is covered:

- Integrate data from one or more sources, which may be from a variety of collection modes, e.g. sampled data, administrative data or open data extracts that have been scraped from the web
- Classify and code the input data
- Review and validate to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding
- Edit and impute to correct any identified problems and impute for non-response to reduce non-response bias
- Derive new variables and units for variables and units that are not explicitly provided in the collection but are needed to deliver the required outputs
- Calculate weights for unit data records according to the methodology created during the design phase
- In the case of sample surveys, weights can be used to 'gross up' results to make them representative of the target population, or to adjust for non-response in total enumerations
- Calculate aggregates and population totals
- This may include summing data, determining measures of average and dispersion, or applying weights to derive appropriate totals. In the case of sample surveys, sampling errors may also be calculated
- Finalise data files in readiness for analysis
- Data are transformed into statistical outputs and includes the production of additional measurements such as indices, trends or seasonally adjusted series
- Analysis is validated in accordance with the Aqua Book[i]; Interpret and explain the outputs by assessing how well the statistics reflect their initial expectations, viewing the statistics from all perspectives using different tools and carrying out in depth statistical analyses
- Apply disclosure control to ensure that the data do not breach the appropriate rules on confidentiality - this may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques
- Finalise outputs to ensure they are fit for purpose and reach the required quality level. This will include: collating supporting information, including interpretation, commentary, technical notes, briefings, measures of uncertainty, etc.
- Ensure the confidentiality of individuals/businesses is protected through the application of appropriate disclosure control techniques

Statistical Knowledge required

Survey Methodology	
• Estimators (Horvitz-Thompson, expansion estimators, ratio estimators, GREG estimators, variance estimators, use of auxiliary information) • Weighting (design weights, weighting for non-response, post stratification, calibration, trimming) • Editing and imputation (detecting and correcting errors, Winsorisation, multiple imputation)	• Small area estimation (design- and model-based estimators, synthetic estimators, borrowing strength over space and time) • Total Survey Error, bias, variance • Minimising response burden • the risk of non-response bias • Maximising response rates and minimising

Descriptive Statistics • Measures of location (different averages, percentiles) • Measures of dispersion and other features of a distribution (interquartile range, skew, kurtosis)	• Measures of uncertainty (standard errors, coefficients of variation, confidence intervals) • Disclosure control
Regression • Multiple linear regression models • Estimation and inference in multiple linear regression • Regression diagnostics (such as leverage and influence, residuals, normality of errors)	• Variable selection techniques • Generalised linear models (link functions, logistic regression, log-linear models, probit models, Poisson regression) • Regression discontinuity designs

Analysis of Variance
• Analysis of variance (ANOVA)
• Multivariate designs (MANOVA)
  • Between-groups design
  • Repeated measures
• Analysis of covariance (ANCOVA/MANCOVA)
	• Testing the assumption of homogeneity of variance sphericity
	• The Kruskal-Wallis test as a non-parametric alternative
	• Post-hoc tests

Multivariate analysis • Principle Components Analysis • Factor Analysis • Discriminant Function Analysis	• Cluster Analysis • Image analysis • Spatial Statistics
Hypothesis Testing • Type I and Type II errors • P-values, significance levels and power calculations • Common parametric tests (e.g. t tests, binomial tests, tests of Pearson's correlation coefficient, tests of regression coefficients)	• Common non-parametric tests (e.g. chi-squared, Mann-Whitney U test, Wilcoxon test) • Correcting for multiple comparisons (e.g. Bonferroni correction)
Time Series • Time series models (autocorrelation, ARIMA processes, state space models and the Kalman filter, fitting and validating models) • Forecasting	• Seasonal adjustment (canonical decomposition, temporary and permanent prior adjustments, Easter effects, working day adjustments and other calendar effects).
Index Numbers • Unweighted indices (Carli, Jevons, Dutot) • Weighted indices (Laspeyres, Paasche, Lowe) • Superlative indices (Fisher, Tornqvist, Walsh)	• Chain linking • Deflators • Hedonic method
Statistical Quality • Understanding the different dimensions of statistical quality • Strategies for quality management	• Measuring and reporting statistical quality • Use of harmonised standards to help drive up quality

Useful Statistical Programming Languages

Languages	*Resources*
• SAS • R • Python • X13ARIMA-SEATS • SPSS • Stata • Blaise • Javascript • MLwiN • Excel and VBA	• See GSS Learning Curriculum for hyperlinks to available courses/e-learning • Journal of Statistical Software • Excel VBA Tutorial • An Introduction to R • Learning SAS by Example: A Programmer's Guide; Cody, R.

Examples

Application across GSS	*Useful Resources*
Survey Methodology **DWP:** Family Resources Survey 2010/11 (methodology chapter) **ONS:** GSS Methodology Series, Sample Design Options for an Integrated Household Survey **Scottish Government:** Scottish Population Surveys Centralised Weighting Project	• Sampling Techniques; Cochrane, W.G. • Survey Sampling; Kish, L. • GSS Survey Methodology Bulletin (archived website) •
Descriptive Statistics **DWP:** Households Below Average Income **UK Census offices :** Disclosure control policies for the 2011 censuses in the UK	• GSS Guidance: Guidance on statistical disclosure control
Regression **DEFRA:** Multiple regression, Demographic patterns in key dietary indicators, Family Food 2013 **DWP:** Linear regression, Training and progression in the labour market **MoJ:** Regression discontinuity design, The effect of early release of prisoners on Home Detention Curfew (HDC) on recidivism	• Introduction to Linear Regression; Lane, M. • An Introduction to Generalised Linear Models, Dodson, A. J. • Practical Regression and Anova using R; Faraway, J. • A Modern Approach to Regression with R; Sheather, S. •
Analysis of Variance **ONS:** One-way ANOVA, UK Time Use Survey	• Handbook of Parametric and Nonparametric Procedures; Shiskin, D.
Multivariate Analysis **Cabinet Office:** Factor analysis, Civil Service People Survey 2014 Technical Guide **DCLG:** Factor analysis in English Indices of Multiple Deprivation 2015 Technical Report (appendix E) **DEFRA:** Principal components analysis, Baseline management and analysis of UK ozone **HSE:** Factor analysis, The effects of transformational on employees' absenteeism **ONS:** k-means cluster analysis, 2011 Census area classifications **ONS:** Estimates using principal components	• Interpreting Multivariate Data; Barnett, V.

analysis, Forecasting GDP using external data sources	
Hypothesis Meeting **ONS:** Improving ONS's Advance Letter for Social Surveys: a Split Sample Trial on the Opinions and Lifestyle Survey (GSS Survey Methodology Bulletin no. 73)	• http://www.ats.ucla.edu/stat/ • Handbook of Parametric and Nonparametric Procedures; Shiskin, D.
Time Series **DECC:** Structural vector auto-regression model, Fossil Fuel Price Projections **ONS:** Time series modelling, Modelling the UK Labour Force Survey using a Structural Time Series Model **Welsh Government:** Time series modelling, Seasonal adjustment and road casualty data	• ONS Guide to Seasonal Adjustment, 'The Black Book' • Forecasting, Structural Time Series Models and the Kalman Filter; Harvey, A. • Time Series Analysis and Its Applications; Shumway, R. H. • Time series: theory and methods; Brockwell, P.
Index Numbers **Defra:** Wild bird populations in the UK **DCLG:** English Indices of Multiple Deprivation 2015, Technical Report **NISRA:** House Price Index, Methodology Note **MOD:** Measuring Defence Inflation	• Price and Quantity Index Numbers; Balk, B. M. • International Labour Organisation, Consumer Price Index Manual: Theory and Practice 2004
Statistical Quality **HSE:** RIDDOR Statistics, Background Quality Report **MoD:** Defence Inflation Statistics, Background Quality Report **NISRA:** ONS report on developing quality measures for Northern Ireland construction statistics	• GSS website: Quality section

3. Presenting and disseminating data effectively

This statistical strand is concerned with the release of the statistical products to customers. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support customers to access and use the outputs released by the Department.

This statistical strand maps onto the **Disseminate and Evaluate phases of the GSBPM**. The following is covered:

- Manages the update of systems where data and metadata are stored for dissemination purposes
- Produces dissemination products to meet user needs - this could include printed publications, press releases, infographic sheets, interactive web sites (graphics), web pages, downloadable files, etc.
- Project manages the release of dissemination products
- Provides briefings for specific groups such as the press or Ministers
- Operates within the arrangements for any pre-release embargoes
- Promotes the dissemination of statistical products to help reach the widest possible audience
- Manages user support to ensure that customer queries and requests for services are recorded, and that responses are provided within agreed deadlines. These queries should be reviewed regularly to provide an input to the over-arching quality management process, as they can indicate new or changing needs.

Statistical Knowledge required

Data Visualisation • Understanding what chart types are most appropriate for depicting different relationships • Static visualisations	• Interactive visualisations • Infographics • Mapping

Communicating Statistics
• Writing about statistics
  • Statistical commentary for non-technical audiences
  • Presentation of official statistics
• Making data meaningful
	• Communicating uncertainty and change
	• Effective use of tables and graphs
	• Releasing statistics in spreadsheet

Examples

Application across the GSS	*Useful Resources*
Data Visualisation **BIS:** Interactive data visualisation tool, International trade in goods **DCMS:** Treemap diagram on page 8 of the Creative Industries Economic Estimates **DWP:** Universal Credit interactive map **Institute for Government:** The Whitehall monitor, Coalition in 163 charts **ONS:** Maps and visualisations, Claimants of Jobseeker's Allowance (JSA)	• GSS guidance: guidance on graphs and tables • Show Me the Numbers – Designing Tables and Graphs to Enlighten; Few, S. • The Visual Display of Quantitative Information; Tufte, E. R.
Communicating Statistics **DEFRA:** Wild Bird Populations in the UK	• GSS guidance: presentation and dissemination

(annotated by Good Practice Team) **DfE:** Implementing a new release format at the Department for Education **DWP:** Universal Credit monthly experimental official statistics and Work Programme National Statistics	

[i] The Aqua Book was introduced by HMTreasury in 2015; the Book provides guidance for all professions on the production of quality analysis for government. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government