Statistical Report on the Effects of Compulsory Age-determined Retirement Policy for Academic Staff

June 18, 2018

To the Employment Tribunal

Summary

This report was compiled at the request of and using data provided by Professor Paul Ewart, Worcester College, University of Oxford; it was specified from the start that the report would be objective and profess no personal opinion concerning any University policy. Professor Ewart provided instructions in the form of 5 specific questions be addressed; these are detailed under the section entitled Instructions which precedes the Introduction and appears on the next page.

The data were obtained by Professor Ewart from the Higher Education Statistics Authority and Annexe F of EJRA: Review Working Group Report published by the University of Oxford.

The report's conclusions are as follows:

- the data provide no evidence that the policy of Employer Justified Retirement Age (EJRA) practised by Oxford and Cambridge has had any effect upon the promotion of gender equality;
- the data provide no evidence that the EJRA policy has had any effect in promoting inter-generational fairness;
- the data provide no evidence that the EJRA policy has produced career opportunities for the younger generation;
- the data provide no evidence that the EJRA policy has resulted in the proportion of Academic Staff over the age of 65 being significantly different from that same proportion in the rest of the Russell Group of Universities;
- by 2017 the proportion of female academic employees, whether on permanent/open-ended or fixed-term/temporary contracts, had fallen away from an initial position of equality with the Russell Group;
- from 2012 onwards the proportion of female employees on permanent/open-ended contracts has shown a downward trend, in contrast with a steady, upward trend in the Russell Group;
- the proportion of female academic employees in Cambridge has risen for the temporary/fixed-term contract holders, but remains well below the level of the Russell Group;
- the proportion of female academic employees in Cambridge with permanent/open-ended contracts has fallen further behind the Russell Group, rising more slowly from 2006−07 and from 2015 showing a trend which is neither increasing nor decreasing.

Instructions

Instructions were given by Professor Ewart in the form of a request for answers to 5 questions, such answers to be obtained from data which he had provided. Appendix 1 contains a statement from Professor Ewart in which the data sources and the format in which they were presented for analysis are specified.

The questions were as follows.

1. Do the data provide any evidence that the EJRA policy at Oxford (and Cambridge) is an effective means of improving diversity in relation to the proportion of women in academic posts?
   - Specifically, is there any evidence that the EJRA policy has led to an improvement in gender diversity (proportion of women) in the Statutory Professor grades in Oxford?
   - Is there any evidence that the policy has led to an improvement in gender diversity (proportion of women) in Associate Professor grades in Oxford?
2. Do the data provide any evidence that the EJRA policy is an effective means of promoting inter-generational fairness in relation to creating more opportunities for young academics to obtain permanent academic posts in Oxford?
3. Do the data provide any evidence of a significant difference in the age at which academics are appointed to posts at Oxford and Cambridge compared to the other Russell Group universities?
4. Do the data provide any evidence that a lack of a compulsory retirement policy at the other Russell Group universities has resulted in a significant increase in the proportion of older academics occupying permanent posts?
5. Is there any evidence of changes in the number of fixed-term or "career development" posts relative to permanent posts and any consequent effect on gender diversity in Oxford?

Introduction

The Employer-Justified Retirement Age (EJRA) policy of the Universities of Oxford and Cambridge is a compulsory retirement policy based on age. In both Oxford and Cambridge retirement is compulsory at the end of the academic year in which the age of 67 is reached; before the beginning of the academic year 2017−18 Oxford extended this to the age of 68 but Cambridge did not. The principal objectives of the policy have been stated to be the promotion of equality and diversity, the promotion of inter-generational fairness and maintaining opportunities for career progression. Within the Russell Group of Universities, the Universities of Oxford and Cambridge are the only ones to implement an age-determined compulsory retirement policy, which began in the academic year 2011-12 and has continued since then, thereby providing data on its effects. 1

In compiling this report two sources of data have been used, one being data held by the Higher Education Statistics Agency (HESA) and the other being taken from Annexe F of EJRA: Review Working Group Report published by the University of Oxford. The data are described in detail in Appendix 1; broadly speaking they comprise gender-specific and age-specific employment figures recorded from the Russell Group of Universities. Although Imperial College, London is included in this group, it is highly atypical in its cross-section of academic disciplines and has been excluded from all analyses because it is clearly unsuitable as a comparator and can only risk introducing bias. There is a lack of data on racial equality and diversity amongst university employees. There are data on gender and age distribution in employment categories over time which are relevant to career progression and inter-generational fairness. 2

In the sections which follow summaries of what the data show are set out when; 3

(i) comparing Oxford and Cambridge on the one hand with the other Russell Group universities (which have no compulsory retirement schemes) on the other;

(ii) comparing the makeup of Oxford's employment categories from 2006−07 to 2015−16, using Oxford's own published figures; these data therefore cover the periods before and after the introduction of the EJRA in 2011−12.

In what follows the presentation of the available statistics is assisted by including explanatory graphics and brief explanations of the standard statistical techniques employed in analysing those data. For the benefit of the expert analyst, technical details of these techniques, along with validations of their applicability, are given in Appendix 2. All analyses were carried out using the statistical package R, which is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. It is freely available and is widely regarded by academic statisticians as the best and most reliable package for serious data analysis and research. 4

Statistical analysis

Comparison of gender proportions in Oxford and Cambridge with those in the Russell Group

As previously stated Imperial College, London, being a Science and Technology university, is not comparable with any of the other universities and has been excluded. This leaves 21 other Russell Group universities; in this report the term "Russell Group" is taken to mean this group of 21. 5

Throughout the analysis great care has been taken to avoid falling foul of what is technically known as Simpson's paradox. This is a phenomenon which can occur when proportions obtained from aggregated data are compared and it can lead to misleading conclusions. A detailed explanation with illustrative examples is provided in Appendix 4, but as far as the main body of this report is concerned it is not an issue and familiarity with it is not required for understanding the statistical results being presented. Appendix 4 also contains a brief explanation of how to interpret the statistician's quantification of weight of evidence as expressed in terms of p-values. Although they are included in the text, a detailed understanding them is not strictly necessary because they have been quoted alongside symbols which represent the strength of the statistical evidence as follows: 6

.	Very weak evidence
$*$	Reasonable evidence
$**$	Strong evidence
$***$	Extremely strong evidence

Technical problems do not arise when gender proportions from aggregated data are compared in terms of changes in gender profile over time for Oxford, Cambridge and the Russell Group; time is expressed as the ending date of the academic year. Clearly the Proportion of Female Academics can be plotted against Year as a sequence of points on a graph, but this is not a useful display for making comparisons with Cambridge and 21 other Russell Group universities and therefore a statistical technique (known as multilevel cubic spline regression) has been used to produce smooth curves representing the way the proportions change over time. This is illustrated in Figure 1, where the points are shown for Oxford from 2007 to 2017 along with the fitted curve. 7

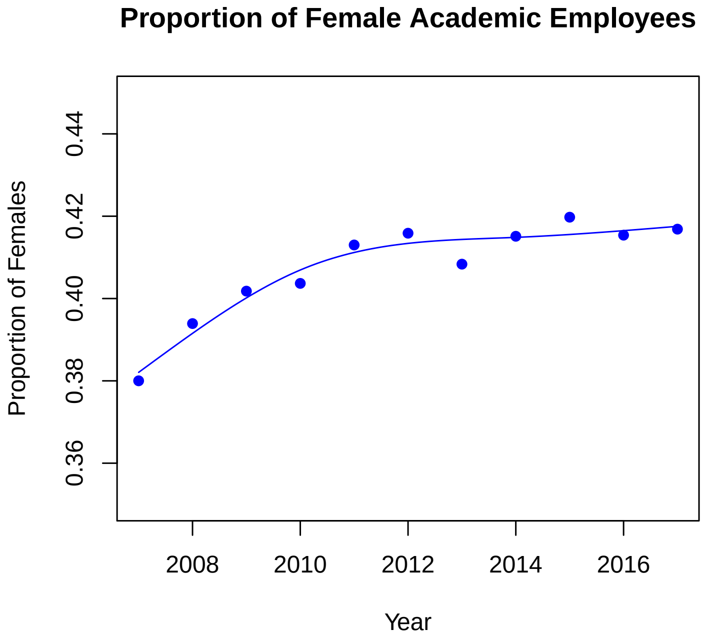


Figure 1: Proportion of Female academic employees in Oxford against Year

Precise details of the statistical methodology behind producing curves such as this, which give an accurate representation of the trend over time, are given in Appendix 3, where validation of the methodology is also shown. In this way Figure 2 was produced, which compares the growth over the period 2007−17 of the total proportion of female academic employees for Oxford, Cambridge and the Russell Group. 8

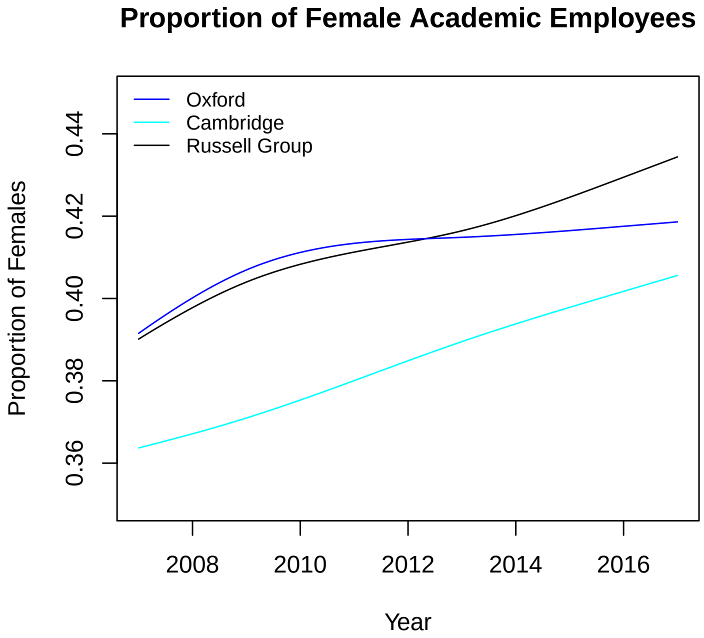


Figure 2: Proportion of Female academic employees against Year

The most noticable features are that both Cambridge and the Russell Group are steadily increasing their female proportions with Cambridge starting at a lower level and with that difference being maintained over time. Oxford, having practically the same proportion of females up to 2012 has subsequently levelled out 9 to a position mid-way between the two. There is no indication that employment policies adopted from 2012 onwards have produced a higher rate of gender-balance improvement in Cambridge and Oxford seems to be levelling off towards no annual change.

The drop from 2012 onwards has resulted in Oxford's proportion of female academic staff becoming significantly lower ($* \; p = 0.011$). Note the single $*$ quantifying the weight of evidence,,which appears here for the first time. Technical details of the statistical test used are given in Appendix 2. 10

However, Figure 2 does not tell the whole story because not all employees are on the same kind of contract. Figure 3 shows the effect of stratifying on type of contract; the dashed lines show the proportions of females on temporary and fixed-term contracts whereas the solid lines show the proportions on permanent, open-ended contracts. 11

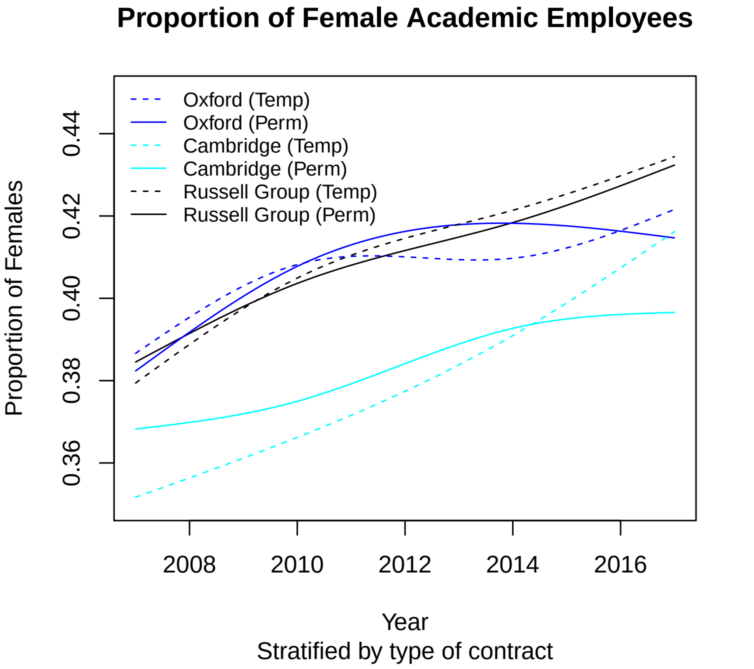


Figure 3: Proportion of Female academic employees against Year

The Russell Group has practically equal gender balance in the temporary and fixed term contracts with both types of contract maintaining a steady increase in the proportion of female staff. In 2007 there is obviously no difference in gender-balance between Oxford and the Russell Group in both temporary and permanent contracts and they remained practically the same until 2012, when the proportion of females with temporary contracts ceased to show an annual rise. By 2014 this had recovered and from then on matched the Russell Group in rate of increase but stayed at a slightly lower level. However, a very different picture emerges for the gender balance among those in Oxford with permanent contracts. From a peak in 2012−13, where it matched the Russell Group, the proportion began to decline and that overall downward trend continued. Between 2013 and 2017 the change in proportion was not large, being of the order of 0.5%, but Oxford was no longer continuing to match the performance of the Russell Group in this respect; furthermore the downward curvature seen in the graph is statistically significant ($* \; p = 0.0125$). 12

In order to look into possible reasons for Oxford's comparative decline in the proportion of female academic staff the data were re-fitted after stratification on Division. The Oxford definition of Division was used to allocate divisional structure to all of the other universities. The results are shown in Figure 4. 13

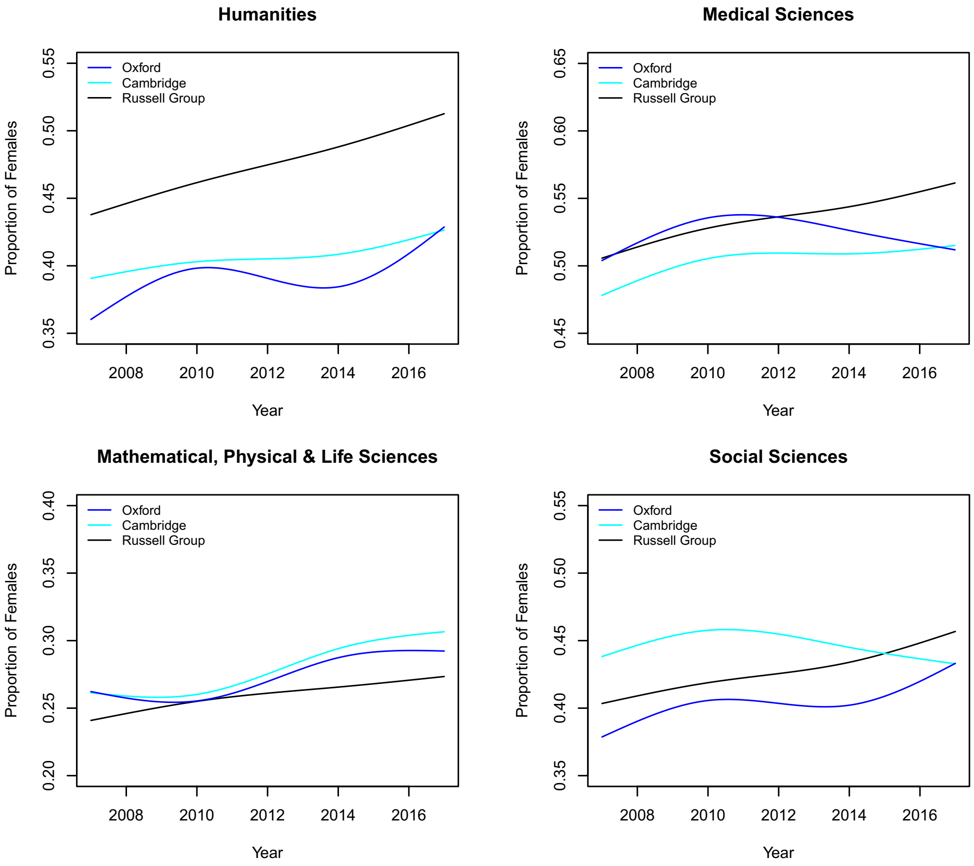


Figure 4: Proportion of Female academic employees against Year stratified by Division (Note different values on the vertical axes of these graphs, but note that the ranges are the same)

The Russell Group universities showed steady rises in the proportion of female academic staff across all Divisions. In the Humanities all of the universities showed the same kind of overall increase; Oxford experienced a downward wobble between 2009 and 2013 but then recovered its earlier upward trend with a rate of increase to match that of the Russell Group. Even so, both Oxford and Cambridge remained behind the Russell Group, being some way short of its achievement of 50%. 14

The most surprising profiles are in the Medical Sciences. The Russell Group continued to increase its Female staff proportion in a steady way, Cambridge showed signs of flattening off and Oxford showed a marked and steady decline which began when a corner was turned after 2010. The difference between the 2010 proportion and the 2017 proportion is not significant (. $p = 0.085$) but even though weak this does constitute evidence. 15

In MPLS the proportions and their time-profiles were almost identical, though there is a slight but insignificant hint of Oxford flattening off from 2015. In the Social Sciences Oxford showed a small downward trend between 2011 and 2013 but from then the position recovered and, overall, matched the rest of the Russell Group whilst remaining slightly lower. Cambridge, however, showed a marked decline in the female proportion with no indication of impending recovery. 16

Nowhere is there any indication that gender-balance is positively affected by the introduction by Oxford and Cambridge of age-related employment policies in 2011−12. 17

Comparison of proportions of females in different staff grades for Oxford, Cambridge and the Russell Group is not possible because their designations in the different universities are not equivalent and, in any case, not available pre-2011. However, Oxford has published its own gender/staff group figures for the period 2006−15, thereby covering the pre- and post-EJRA periods, and these were used to assess the effects of EJRA in the next sub-section. 18

Comparison of gender proportions across staff grades in Oxford

19 Of necessity this part of the analysis is confined to Oxford because its particular staff grades do not apply in other universities; the four grades Statutory Professor, Associate Professor, RS IV, Academic and Academic-Related were considered. The staff grade RS IV is peculiar to Oxford. It is awarded without advertisement and releases academics from normal duties to concentrate on research. The data for this subsection were taken from Annexe F of EJRA: Review Working Group Report published by the University of Oxford and the staff grades are therefore not subject to HESA's interpretation of equivalent classifications. These data are given in the table below; note that in its publication Oxford labelled the academic year in terms of it starting date rather than the end date used throughout the rest of this report, and therefore the year in the left-hand column has been increased by 1 year for consistency.

	Statutory Prof.		RS IV		Assoc. Prof.		Acad. & Acad. Rel.	
Year	Female	Male	Female	Male	Female	Male	Female	Male
2007	23	204	15	37	285	878	1954	2313
2008	21	218	18	47	272	835	2270	2537
2009	25	222	22	68	287	843	2515	2681
2010	23	212	29	79	280	829	2807	2909
2011	24	210	29	82	290	831	2947	3034
2012	27	216	31	84	289	798	3030	3082
2013	29	222	35	102	285	786	3251	3283
2014	27	224	39	125	308	810	3589	3501
2015	30	230	47	148	325	851	3887	3833
2016	34	214	51	157	330	855	4167	4070

Figure 5 is a visual representation of these data and shows the time profiles for the proportion of female staff in the four Oxford grades.

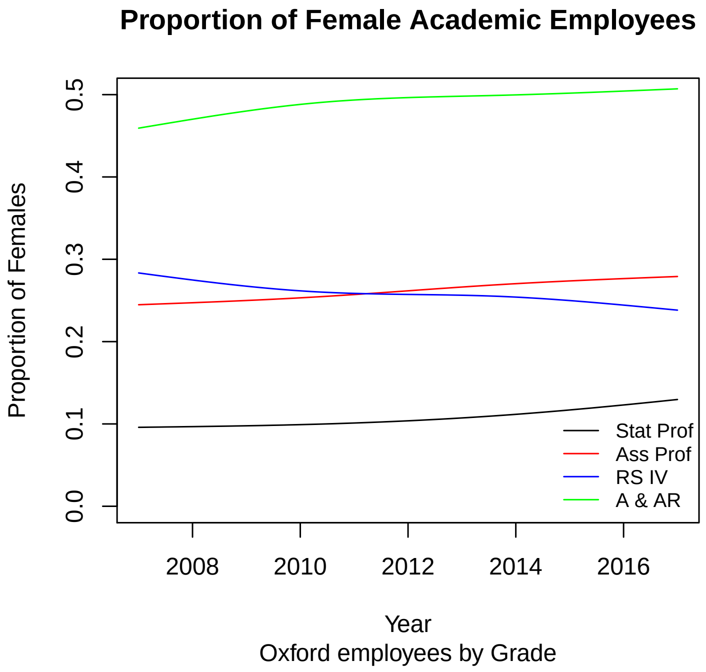


Figure 5: Proportion of Female academic employees against Year by Grade

20 There has been very little change. The proportion of female academic staff in Grade RS IV seems to have been declining very slowly whereas the proportions in the other three grades have shown very slow, steady increases.

21 Figure 5 suggests fitting separate linear relationships to the four grades for each of the two time periods

2007 to 2012 and 2012 to 2016 and testing each pair of fitted lines for different slopes; this would formally test whether there was any evidence of change following the introduction of the EJRA. However, with four simultaneous tests being conducted any assessments of the strength of evidence in the form of p-values obtained is weakened by the fact that four attempts have been made. Clearly this must be so; if a horse were to be backed at 20:1 one would be surprised if it were to win, but if 20 different horses were backed at 20:1 in 20 different races, a win for at least one of them might reasonably be expected. The same is true when interpreting simultaneous statistical tests. With tests being carried out involving 4 staff grades each of the p-values obtained needs to be multiplied by a factor of 4 for interpretation of the weight of evidence it conveys. This is technically known as a *Bonferroni* correction and is applicable here.

Statutory Professor

The estimated growth rate was a multiplicative annual compound growth of 3.8% per annum; formal testing resulted in this not being significantly different from zero ($p = 0.0670$, Bonferroni corrected to 0.2680). The change in growth-rate after 2011 was estimated to be practically zero ($p = 0.9974$) and there was no evidence of any change post 2011. 22

An assessment of the growth-rate of the proportion of females in statutory professorships in the Russell Group was obtained for the period 2013 to 2017; this was a compound 4.5% per annum. Whilst this is significantly greater than the Oxford growth-rate, too much should not be read into this as the HESA definition of the Russell Group posts may not correspond and the estimates were obtained over different time periods. 23

Associate Professor

Here there was very strong evidence of a steady growth in the Female/Male ratio ($*** \; p = 0.0001$, Bonferroni corrected to 0.0004), and again no evidence of any change post 2011 ($p = 0.9022$). The annual growth rate was a compound 2.1% growth per annum which remained unchanged from 2011 to 2016. 24

RS IV

There was no evidence of any annual growth in the Female/Male ratio ($p = 0.4656$) and no evidence of change post 2011 ($p = 0.3815$). 25

Academic & Academic Related

There was evidence of steady annual growth ($* \; p = 0.0028$, Bonferroni corrected to 0.0112) but no evidence of a change post 2011 ($p = 0.2929$). There was compound annual growth ot 2.3% per annum which remained unchanged from 2006 to 2016. 26

Overall summary for the four grades

With the exception of Statutory Professor and RS IV, there was statistical evidence of annual growth in the Female/Male ratio and the annual growth rates did not change after 2011 in any grade. 27

For simultaneous inference from four separate staff grades, Bonferroni corrections should be and were applied to the p-values of the estimated annual growth-rates. Thus in the Statutory Professor grade, although the growth-rate was higher than the others and gave a compound 3.8% growth per annum, with a Bonferroni-corrected p-value of 0.27 it is not close to being statistically significant from zero; in fact, even without the correction it would not have been statistically significant. It should, however, be noted that, with comparatively low numbers of Statutory Professors, a non-significant result is hardly surprising and should not be regarded as cause for dismissing altogether the presence of growth. 28

However, the main result lies in there being no significant change detected in any of the growth rates post 2011. For Statutory Professors and Associate Professors the estimated change in rate is, in fact, zero (to four decimal places) and, although RS IV and Academic & Academic Related come out as slightly negative when taken to that many decimal places, the smallest uncorrected p-value is 0.2929 and there is no evidence for considering any change brought about by the EJRA to be anything other than zero. 29

Of all the grades considered above the fastest annual growth-rate in the Female/Male ratio was 3.8% which 30 was less than the annual growth-rate of 4.5% recorded by the Russell Group professorial grade, the Oxford equivalent of which encompasses a subset of the combined Oxford grades.

Comparison of age distributions in Oxford and Cambridge with those in the Russell Group

The data considered in this section were provided by HESA and comprised the date of appointment, age at appointment, leaving date and age at leaving of every individual academic employee in Oxford, Cambridge and the rest of the Russell Group. From these data the age of every member of academic staff was calculated for the academic years of interest. As before data from Imperial College were removed prior to analysis. Figures for the years 2012−13 and 2016−17 are analysed in the next sub-section; the total numbers of academic staff in 2012−13 were Oxford: 2384, Cambridge: 2311, Russell Group: 36570. In 2016−17 the totals were Oxford: 2821, Cambridge: 2383, Russell Group: 41550. 31

In this section there are several graphs illustrating age distributions of academic staff. It should be noted that these are not displayed in the form of histograms, which have fallen out of favour with many statisticians because of their susceptibility to changing shape if their vertical bars are slightly shifted sideways; this is technically known as bin-edge effect and, as a result, so-called density traces are preferred as the distributional shapes they produce are robust and reliably accurate. The density traces which follow are to be interpreted in terms of area beneath the curve. Thus in Figure 6, which follows immediately below, the area beneath the curve between say, age 40 and age 50, represents the proportion of academic staff between those ages. 32

Age distributions of all academic staff

Figure 6 shows the density traces of the age distributions for Oxford, Cambridge and the Russell Group. 33

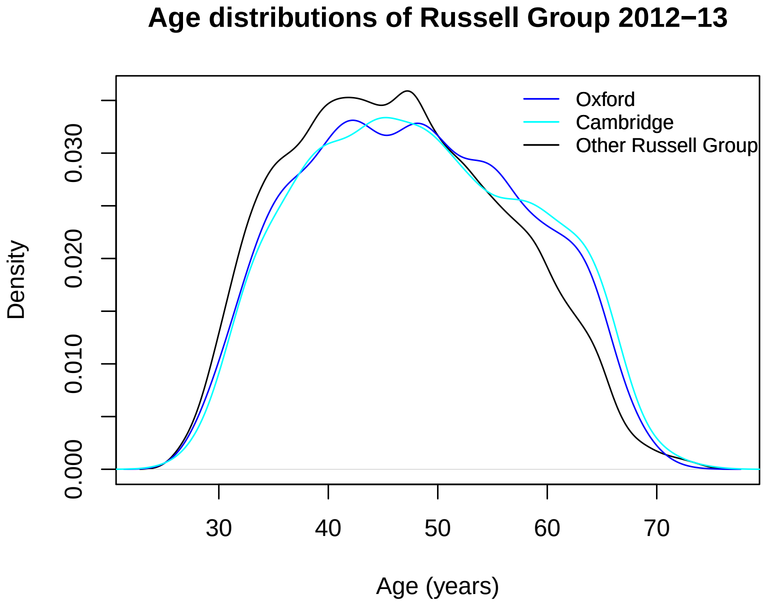


Figure 6: Empirical density traces for academic year 2012−2013

Oxford and Cambridge virtually overlap, whereas the Russell Group distribution lies a little to the left. All of the densities show some right-skew and therefore their medians were used for formal testing. The median age for both Oxford and Cambridge is 48 and for the Russell Group is younger at 46. Formal tests of equality resulted in no significant difference between Oxford and Cambridge ($p = 0.1697$) and highly significant differences of both from the Russell Group ($* * *\ p < 0.0001$). There are no detectable differences in the extreme upper tails of the distributions.

Figure 7 shows the equivalent densities for 2016-17. 34

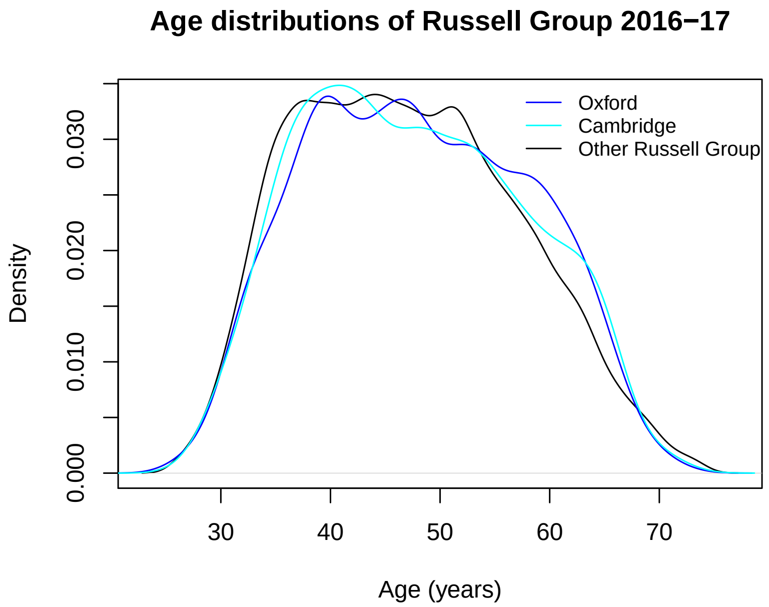


Figure 7: Empirical density traces for academic year 2016-17

Although the shapes are a little different, there has been no change. Formal testing of the age differences again has Oxford and Cambridge staff being significantly older than those in the Russell Group ($\ast\ast\ast$ Oxford: $p < 0.0001$, Cambridge: $\ast\ast$ $p = 0.0018$) with no significant difference between the two of them ($p = 0.2282$).

But, in view of their EJRA policies, the main interest lies in the extreme upper-age tails which the graph 35
in Figure 7 indicates to be indistinguishable from each other. By 2016-17 the EJRA has become established in Oxford and Cambridge and therefore the area under the extreme over-65 tail (*i.e.* the proportion of over-65s) for those universities which do not operate age-related compulsory retirement could be expected to be significantly greater than the areas for those who do. In the Russell Group the proportion is 3.52%, for Oxford it is 3.01% and for Cambridge it is 3.31%. Formal statistical tests for different proportions show that the difference between the Russell Group and Oxford is not statistically significant ($p = 0.1679$) and is not significant for Cambridge ($p = 0.4226$). This means that, for the academic year 2016-17 at least, the Russell Group's having retirement as a matter of personal choice has not resulted in an inordinately high population of academic staff over-the age of 65 and has not resulted in their being any different from Oxford and Cambridge in this respect.

Growth-rates in numbers of Academic Staff

The lack of difference in the proportions of over-65s between the universities cannot be attributed to their 36
having different annual growth-rates in their total academic staff. Annual growth-rates of staff totals were calculated and that of the Russell Group formally tested against the other two for differences, resulting in a p-value of 0.86. It was therefore clear that there was no evidence to suggest that the growth-rates were different. The Russell Group and both Oxford and Cambridge have all been growing at a steady, significant, compound rate of 3.7% per annum ($\ast\ast\ast$ p-value < 0.0001).

Age at appointment

Figure 8 shows the distributions of ages at appointment of those newly appointed in the different universities 37
during the academic years 2012−13 and 2016−17. In 2012−13 Oxford made 46 new appointments, Cambridge made 66 and the Russell group of 21 universities made 1068. In 2016−17 the figures were Oxford: 37, Cambridge: 68 and the Russell Group: 858.

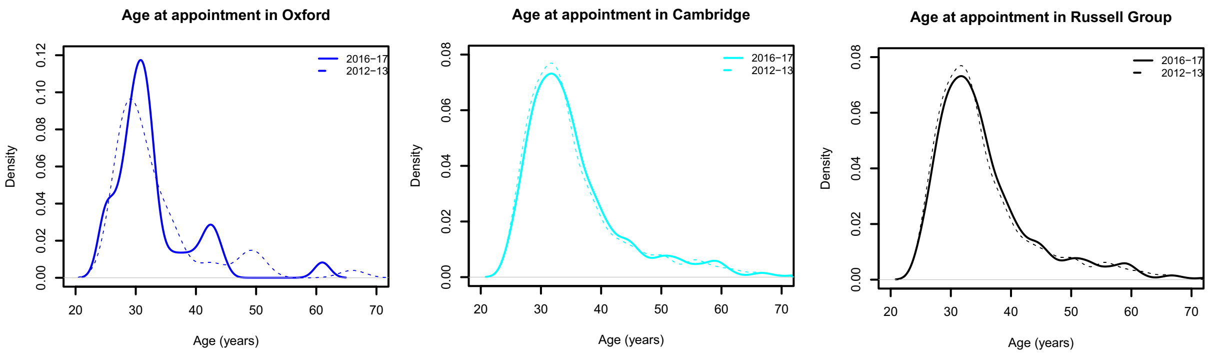


Figure 8: Age at appointment

It seems clear, at least in terms of the modal positions (*i.e.* the positions of the peaks), that there has been little or no change from 2012-13 t0 2016-17 in Cambridge and the Russell Group and their distributions are almost identical; Oxford's mode has moved slightly towards older appointments to bring the mode of the distribution into line with the others. But appearances can be deceptive and formal testing of location shows a different picture. Figure 9 shows comparative boxplots of the data. 38

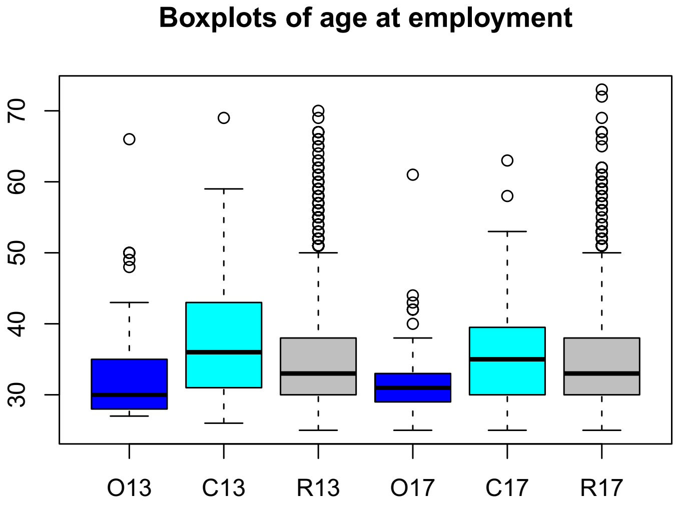


Figure 9: Comparative boxplots of employment age data

A boxplot represents the middle 50% of the data by the edges of a rectangle with the median marked inside it as a line. Whiskers (the technical term for the T-pieces growing from the ends of the box) stretch as far as the furthest data point within 1.5 box-lengths on either side and points outside this range are represented by small circles. Thus one can obtain an intuitive feel for the range and distribution of the data and visually compare different groups. 39

These boxplots show that the median age at employment in Oxford has risen and become a little less concentrated but remains below that of the others, whereas at Cambridge it has reduced very slightly towards that of the Russell Group; the boxplots confirm the stability of the Russell Group's age distribution. 40

All of the boxplots confirm the heavy right-hand skewness (*i.e.* towards the higher ages) of all of the distributions so non-parametric testing of equality of the median ages at appointment was used. In 2012-13 Oxford's median employment age was, after Bonferroni correction, significantly lower than Cambridge ($** p = 0.0012$) but not significantly lower than the Russell Group ($p = 0.0693$) whereas Cambridge was higher than the Russell Group ($* p = 0.0134$). In 2016-17 Oxford's age at employment remained lower than that of Cambridge ($* p = 0.0199$) but had moved to being significantly lower than the Russell Group ($* p = 0.0266$); there was no difference between Cambridge and the Russell Group ($p = 0.6612$). Within Oxford comparing the $2012-13$ and $2016-17$ ages at appointment showed no significant difference (p-value $= 0.62$). 41

Part of the overall question of whether the EJRA has lowered the age at employment is whether it has 42 resulted in Associate Professors being appointed at a lower age. Without data being supplied by either Oxford, Cambridge or both the question cannot be answered, but it *is* possible to use Oxford's published data to compare the distributions of proportions across the different age groups for different years. In other words, the question "Has the distribution across age groups changed between 2007 and 2016?" can be answered.

43 There are two possible interesting comparisons depending on whether or not the above-age 67s are to be included as a separate group in the age profiles. The numbers of Associate Professors by age group are:

Age group	*Year 2007*	*Year 2016*	*Age group*	Year 2007	*Year 2016*
Under 30	15	14	Under 30	15	14
30 − 39	238	445	30 − 39	238	445
40 − 49	244	486	40 − 49	244	486
50 − 59	204	387	50 − 59	204	387
60 − 67	61	132	60 +	63	149

Excluding the over 67s a chi-squared test of independence of Age Group and Year has a p-value of 0.2999, so there is no evidence of dependence; in other words the relative proportions recorded in 2007 have not changed by 2016. This conclusion is not changed when the over 67s are included, the p-value being 0.1857. Whichever data are included, there is no evidence of any change in the distribution of relative proportions across the Age groups from 2007 to 2016.

Conclusion

44 In terms of promoting gender equality there is no evidence that the introduction of the EJRA has had the slightest effect. Oxford has, since the introduction of the EJRA, shown a widening of the gap in gender balance with the rest of the Russell Group but Cambridge has not shown this; neither does it seem to be closing the gap. The downward trend in the Medical Sciences, which is not seen with Cambridge, is evident.

45 The Russell Group has shown a steady rise in gender equality in terms of both Temporary/Fixed-term contracts and Permanent/Open-ended contracts; the proportions of female staff for the two types of contract are practically the same and these have risen steadily, in parallel, over the period 2007 to 2017. In Oxford both of these proportions have dropped from positions of equality with the Russell Group in 2012. Following a fall between 2012 and 2014 the proportion on Temporary/Fixed-term contracts has resumed a steady annual increase in parallel with but below the level of the Russell Group. The proportion with Permanent/Open-ended contracts began to show a downward trend from 2013 and that trend has continued through to 2017.

46 Temporary/Fixed-term contracts in Cambridge had a gender balance a long way below that of Oxford and the Russell Group in 2007; this has risen steadily and by 2017 the gap has become narrower although still below the Russell Group level. With Permanent/Open-ended contracts the gap has shown no sign of closing and from 2015 the annual rise levelled off, leaving the gender balance well below both Oxford and the Russell Group.

47 In terms of promoting inter-generational fairness and career opportunities there is no evidence that the introduction of EJRA has had the slightest effect. When it comes to opportunities for the younger generation, age at employment has not decreased and both Oxford and Cambridge closely parallel the rest of the Russell Group. In Oxford the distribution of ages in the Associate Professor group in 2016-17 is very much as it was in 2006-7.

48 Oxford, Cambridge and the Russell Group have percentages in the over-65 age group which are practically the same and certainly not significantly different; thus there is no indication that the Russell Group's lack of an age-related retirement policy has had anything other than a negligible effect on retirement from its universities.

Appendix 1: Description of the data (Professor Paul Ewart)

The data used to prepare this report were provided by Professor Paul Ewart in the data file HESA 55768_data.csv together with the accompanying file HESA 55768_Notes_and_Labels. The data file was provided by HESA under contract: "Quote 55768 Paul Ewart" (see attachment).

The data file contains information on each academic at Russell Group universities in the period covering academic years 2006/07 to 2016/17. Specifically the file contained the following data:

Number of academic staff at Russell Group Higher Education providers 2006/07-2016/17 by HE provider, HEP:

- Contract levels (2012/13-2016/17)
- Sex
- Age (full)
- Date appointed at current HEP (MM/YYYY)
- Date left HEP (MM/YYYY)
- Academic contracts of leavers (ZACLEAV02) (2006/07-2015/16)
- Leaving destination of leavers on academic contracts (grouped) (ZDESTGP01) (2006/07-2015/16)
- Reason for end of contract* (2012/13-2016/17)
- Cost centre (2012/13-2016/17)
- Cost centre (2006/07-2011/12)
- Academic discipline 1 (2006/07-2007/08)∗∗
- Academic discipline 2 (2006/07-2007/08)∗∗
- Academic discipline 1 (2008/09-2011/12)∗∗
- Academic discipline 2 (2008/09-2011/12)∗∗
- Current academic discipline 1 (2012/13-2016/17)∗∗
- Current academic discipline 2 (2012/13-2016/17)∗∗
- Current academic discipline 3 (2014/15-2016/17)∗∗
- Nationality (UK/Other EU/Other EEA/Other Non-EU/Not known)

∗ Restricted only to those with an end date of contract.
∗∗ Based on subject area

The data file consists of a table comprising 20 columns and 827,734 rows. The columns indicated information on the 18 items listed above plus columns indicating the HEP and academic year. This data file was later supplemented by file "55768_variation_data" giving information on "Terms of employment" i.e. permanent/open-ended or temporary/fixed-term contracts.

An additional column of data giving the "Age at appointment" i.e. the age at which an academic was appointed to the current HEP was created as follows. The data giving "age" in a given year were used to calculate the year in which the academic was born, Birth Year. Together with the data giving the "Date appointed at current HEP" a column was created giving the "Age at appointment" at the current HEP.

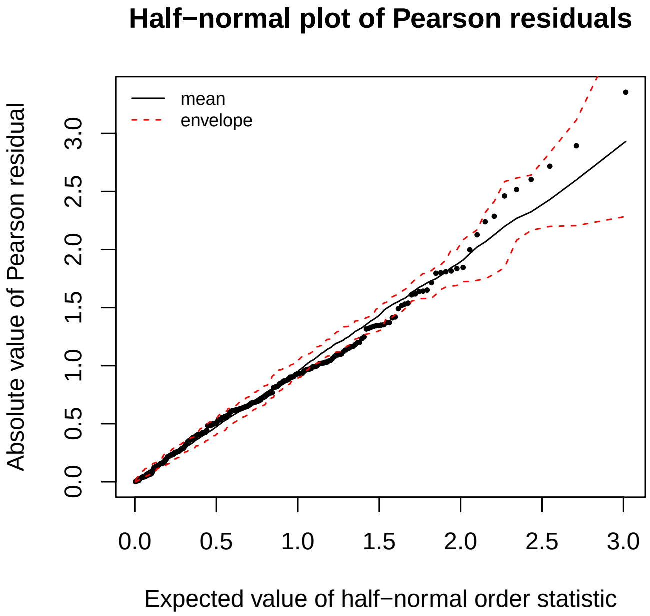
The data file therefore contained over 20 million individual items of information. In order to carry out a statistical study, sub-sets of these data relevant to specific questions were extracted using the program EXCEL as .csv files by Professor Ewart and provided for presentation.

The data presented in Figures 1 and 2 of the report used the sub-set giving the numbers of male and female academics in each HEP in each academic year separated into those in Oxford, Cambridge and the other Russell Group universities excluding Imperial College.

The data presented in Figure 3 consisted of the same data but with the additional information on terms of employment used to separate the data according to the type of contract; permanent/open ended or temporary/fixed-term. In a similar way the data displayed in figure 4 were obtained from the database by disaggregating according to academic discipline using the HESA classifications given in the "Notes and Labels file" corresponding to the 4 Oxford Divisions. Figure 5, as explained in the report, was constructed using the data supplied by the Review Working Group report annexe F. The remaining figures, 6−9 were constructed using the age information contained in the HESA data file and the "Age at appointment" values calculated as explained above.

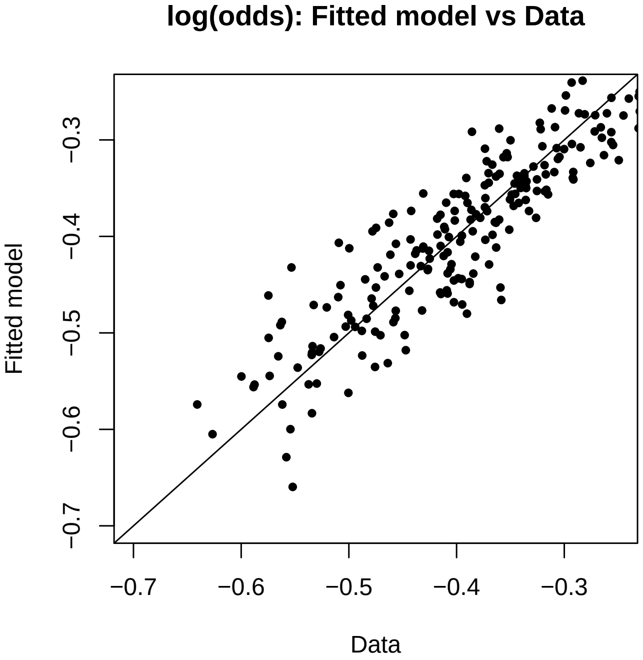
Appendix 2: Technical details

The data for all 23 universities, each of which has its own profile over time for the proportion of its female staff, are multilevel in nature and must therefore be analysed as such. Therefore, in order to cater for separate profiles within individual universities and extract the underlying common trends, multilevel Generalised Linear Models (GLMs) were fitted to the years 2007−2017 with the relationship between Female proportion and Year being fitted by a cubic spline with 3 degrees of freedom. Specifically, the R package *glmmPQL* from the *MASS* library was used to fit a binomial mixed effects model with *Academic Institution ID* as a random-effect level. Statistical models fitted in this way were used to produce Figures 1 to 4. Compliance with the modelling assumptions and the quality of fit were assessed by producing half-normal QQ-plots of the Pearson residuals with 95% envelopes, as recommended by Collett, D. (2003) *Modelling binary data.* Second edition. Chapman & Hall/CRC. The multilevel version of such a plot resulting from the model used to generate Figure 2 is shown below as an example.



As required, the residuals lie close to the solid line and within the envelope, so the model fit is good with no potential outliers. Similar plots were produced for all of the models in this section and all were equally satisfactory.

In order to give a "feel" for the quality of the fit, a further plot was also produced showing the fitted model versus the raw data. The plot, which appears below, shows clearly that the fit is very good.



Again, such plots were produced as a check on all of the models used for comparing gender proportions across the Russell Group plus Oxford and Cambridge.

The hypothesis test referred to in paragraph 10 was a Fisher's exact test of equal proportion. The result should, however, be treated with some caution because the possibility of Simpson's paradox cannot be entirely ruled out; the quoted p-value would only be trustworthy if the relative proportions of the staff totals across the different academic disciplines were to be similar in Oxford, Cambridge and the Russell Group and data for checking this are not readily available. Having said that, given that, without Imperial College, most Russell Group universities have fairly similar ranges of subjects and faculties so the test cannot be excessively misleading.

The treatment of gender proportions across staff grades in Oxford posed different problems. With comparatively small numbers and in order to cater for lack of independence of year-by-year proportions obtained fro them a log-odds growth model was fitted to the years $2006-$ 2015 with the response being the log of the Female/Male ratio for each category. An indicator variable (EJRA) was included as an interaction, thereby providing a switch to detect any change in growth rate from 2012. The models fitted well and satisfied all of the diagnostic tests (see below)

As before, preliminary models were fitted by using a cubic spline and, for these data, both with and without interaction with the EJRA variable. A likelihood-ratio test of the models both with and without this variable showed no significant effect ($p = 0.7657$). The graphs displayed in Figure 5 were obtained from this model.

Formal Shapiro-Wilk tests for normality of the model residuals were carried out with resulting p-values for Statutory Professor ($p = 0.3401$), Associate Professor ($p = 0.466$), RSIV ($p = 0.0999$), Academic & Academic Related ($p = 0.459$); the model residuals showed no evidence of heteroscedasticity or serial correlation and therefore there was no reason to doubt the adequacy of the fits obtained. It should be noted that, with four separate models being considered, Bonferroni corrections were applied in the discussions. The model outputs are given below.

Statutory Professor

The fitted model is shown in the table below.

Coefficient	*Value*	*Std.Error*	*t-value*	*p-value*
(Intercept)	−77.89	34.94	−2.2292	0.0610
Year	0.038	0.017	2.1664	0.0670
Year:EJRA	0.000	0.000	0.0034	0.9974

There is no evidence of a steady growth in the Female/Male ratio ($p = 0.0670$, Bonferroni 0.2680), and there is no evidence to support a change in growth-rate post 2011 ($p = 0.9974$). The Year coefficient for log(Female/Male) was 0.038, giving a multiplicative annual growth of exp(0.038) = 1.038 or a compound 3.8% growth per annum.

Associate Professor

Coefficient	*Value*	*Std.Error*	*t-value*	*p-value*
(Intercept)	−42.67	5.567	−7.6649	0.0001
Year	0.021	0.003	7.4708	0.0001
Year:EJRA	0.000	0.000	0.1274	0.9022

There is very strong evidence of a steady growth in the Female/Male ratio ($p = 0.0001$, Bonferroni 0.0004), and again no evidence of any change post 2011 ($p = 0.9022$). The annual growth rate of log(Female/Male) was 0.021 giving a multiplicative annual growth of exp(0.0211) = 1.021 or a compound 2.1% growth per annum.

RS IV

Coefficient	*Value*	*Std.Error*	*t-value*	*p-value*
(Intercept)	19.25	26.265	0.7330	0.4874
Year	−0.010	0.013	−0.7716	0.4656
Year:EJRA	−0.000	0.000	−0.9338	0.3815

There is no evidence of any annual growth in Female/Male ratio ($p = 0.4656$) and no evidence of change post 2011 ($p = 0.3815$).

Academic & Academic Related

Coefficient	*Value*	*Std.Error*	*t-value*	*p-value*
(Intercept)	−45.66	10.136	−4.5050	0.0028
Year	0.023	0.005	4.4987	0.0028
Year:EJRA	−0.000	0.000	−1.1373	0.2929

There is evidence of steady annual growth ($* \; p = 0.0028$, Bonferroni 0.0112) but no evidence of a change post 2011 ($p = 0.2929$). There is multiplicative growth of exp(0.023) = 1.023 or a compound annual growth of 2.3%.

Figures 6 to 8 were produced using kernel density estimates (Sheather, S. J. and Jones M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* **B**, 683−690). Median differences in this section were tested with Mann-Whitney-Wilcoxon tests and equality of the tail proportions were tested with Fisher's exact test.

Growth of the total number of staff over time was fitted using the *lme* function from the R library *name* with *Academic Institution ID* as a random effect level. The model was fitted with and without interaction terms for the factors Oxford, Cambridge and Russell Group, thereby testing for contrasts in the rate of growth parameters. A likelihood ratio test of the models with and without the interaction produced a p-value of 0.86, thereby showing no evidence of any difference in growth rates between the three categories.

Owing to heavy skewness, non-parametric tests were used to assess differences in age at appointment.

Standard chi-squared tests were used for the tables paragraph 43. With no cell in the tables having an expected value of less than 5, there was no reason to doubt p-values obtained from the asymptotic chi-squared distribution of the Pearson statistic.

Appendix 3

Simpson's paradox

There is a nice example which has the advantage of being from a genuine data set (1986 C. R. Charig; D. R. Webb; S. R. Payne; J. E. Wickham) and was influential in recommending the best technique for the surgical treatment of renal calculi (*i.e.* kidney stones).

There were two basic kinds of operation in use, these being all open surgery (A) and *percutaneous nephrolithotomy* (a small puncture keyhole surgery) (B), the latter having the advantage of quicker recovery time. The following data wer gatherted from a clinical trial:

	Operation A			Operation B		
Stone size	Success	Total	% success	Success	Total	% success
Small	81	87	93%	234	270	87%
Large	192	263	73%	55	80	69%
Totals	273	350	78%	289	350	83%

Now the paradox is this: taken overall (i.e. looking at the line marked Totals) the success rate of A is 78% and of B is 83%, so can one conclude that Operation B is to be preferred? Clearly that cannot be true because, for Small stones, A out-performs B by 93% to 87% and, for Large stones, A out-performs B by 73% to 69%.; therefore A is clearly better than B for both stone sizes.

So what is going on here? The explanation is that surgeons were making pre-operation assessments of stone size and, when anticipating a small stone, were tending to opt for B and were opting for A when a large stone was diagnosed. This led to a large disparity in the operation totals: for Smalls there were 87 As and 270 Bs, for Larges there were 263 As and only 80 Bs; in statistical jargon the surgeons' initial diagnoses and subsequent choices of operation are known as a *confounder*. The incorrect conclusion reached by looking at totals is said to have been *confounded* by the hidden variable of diagnosis.

Another example, which is pertinent here, is that of graduate admissions to Berkeley in 1973, where there was much criticism of the disparity between male and female graduate admissions. The table below gives the figures for the 6 largest faculties.

Gender	Accept	Reject	% Accepted
Male	1198	1493	44.5%
Female	557	1278	30.4%

Clearly there is a bias against accepting females - or is there?

Now suppose we break it down by faculty.

	Male			Female		
Faculty	Accept	Reject	% Accepted	Accept	Reject	% Accepted
A	512	313	62%	89	19	82%
B	353	207	63%	17	8	68%
C	120	205	37%	202	391	34%
D	138	279	33%	131	244	35%
E	53	138	28%	94	299	24%
F	22	351	6%	24	317	7%

In only two out of the six faculties are there higher percentages of male acceptances. What is going on here? The confounder is that admission to some of the faculties is much harder; in other words places are more readily available in some faculties than in others. Suppose we look at the same data in a different way by taking the percentages of all applicants admitted, faculty by faculty, and seeing the male-female breakdowns of those who applied.

Faculty	% Admitted	Male applicants	Female applicants
A	64.5%	825	108
B	63%	560	25
C	35%	325	593
D	34%	417	375
E	25%	191	393
F	6.5%	373	341

Faculty A has the highest admission rate and 8 times as many males as females apply. The second highest is Faculty B with over 22 times as many male applicants as female applicants. The one faculty with more female applicants than males admits only 34%. The only correct way to analyse these data is to take the confounder into account and stratify the analysis across the different faculties. Simply looking at totals leads to the incorrect conclusion of anti-female bias when, in fact, the evidence is to the contrary.

Clearly there are pitfalls involved in looking at data of this kind and in seeking figures from any data source it is essential to guard against being given pre-processed overall percentages and pre-processed totals. Overall figures for looking at universities will not do - they must be broken down by faculty or, if that is not possible, at least by division.

The correct analysis for the Berkeley graduate admissions involves fitting a binomial generalised linear model; it turns out that there is no significant difference in the chances of admission between the genders except in the case of Faculty A, where a female is 2.86 times more likely to be admitted as a male. This is highly significant with $p = 0.0001$ and is the opposite of the conclusion reached from the aggregated figures.

Interpreting p-values

In statistical data analysis it is usual to formulate a question associated with the data in terms of a hypothesis. In particular, one has a so-called *null hypothesis* which refers to some basic premise to be adhered to unless evidence from the data causes it to be untenable. For example, in a clinical treatment data may be collected to compare two treatments (say, old v. new). The *null hypothesis* would then be no difference between treatments and evidence from the date would be looked for with a view to rejecting it and concluding that there was significant evidence of a difference. The weight of evidence is expressed in terms of a *p-value*, which is interpreted as the probability that the observed difference is due to chance rather thanking due to a genuine difference. Clearly a p-value is expressed on a scale of 0 to 1 and the smaller the p-value the greater the weight of evidence.

The obvious question arising is *what is the critical level for the p-value to result in rejection? Is there some generally accepted level at which null hypotheses are automatically rejected?* Alas, the literature is filled with what purports to be the definitive answer to this question, which is so misleading and ridiculous that it needs special mention.

A significance level of $p < 0.05$ is often taken to be of definitive, because it is below the "magic" level of 0.05**.** For example suppose that we had tested a new drug (new drug versus standard drug), which under the null hypothesis of no difference between the two drugs, gave $p = 0.04$. This says that the apparent difference between the two drugs being due to chance is less than 1 in 20. The p-value of 0.05 is the watershed used by the American control board (the FDA, which stands for Food and Drugs Administration) which licences new drugs from pharmaceutical companies. As a result it has been almost universally accepted right across the board in all walks of life.

However this level can be, to say the least, inappropriate and possibly even catastrophic. Suppose, for example, we were considering test data for safety critical software for a nuclear power station, N representing the number of faults detected in the first 10 years. Would we be happy with a p-value on trials which suggests that the probability that N is greater than zero is 0.05? We might be more comfortable if $p = 0.0001$, but even then, given the number of power stations (over 1000 in Europe alone) we would be justified in worrying. The significance level which should be used in deciding whether or not to reject a null hypothesis ought to depend entirely on the question being asked; it quite properly should depend upon the consequences of being wrong. At the very least we should qualify our rejection with something like the following.

.	$0.05 < p \leq 0.06$	“Weak evidence for rejection”
*	$0.03 < p \leq 0.05$	“Reasonable evidence for rejection”
	$0.01 < p \leq 0.03$	“Good evidence for rejection”
**	$0.005 < p \leq 0.01$	“Strong evidence for rejection”
	$0.001 < p \leq 0.005$	“Very strong evidence for rejection”
* * *	$0.0005 < p \leq 0.001$	“Extremely strong evidence for rejection”
	$p \leq 0.0005$	“Overwhelming evidence for rejection”

The asterisks have been put alongside the levels referred to in the main text of the report so that the reader can judge the weight of evidence quickly without having refer to this appendix for checking the quoted p-values themselves.

Statement from the author to the Employment Tribunal

1. I understand my duties to the Tribunal and I have complied with those duties.
2. I have taken into account all relevant material facts and have identified the material on which the report is based.
3. I understand my duty to raise without delay any matter which causes me to alter the contents of this report.
4. I confirm that I have made clear which facts and matters referred to in this report are within my own knowledge and which are not. Those that are within my own knowledge I confirm to be true. The opinions I have expressed represent my true and complete professional opinions on the matters to which they refer.

Since retiring from my Tutorial Fellowship in Mathematics I have given my services, free of charge, to any member of the Academic Staff or Academic Administration requiring statistical advice and/or expertise. I operate a full-time Statistical Consultancy Service within the Department of Statistics, a department which is recognised as being one of the leading statistics departments in the world and which was rated the UK's top department in the UK REF. This Service is restricted to members of Oxford University and is available to all researchers across the University, including DPhil students, and for this I require and receive no salary. My sole objective is to improve the quality of statistical analyses in research papers published by members of the University and to improve the techniques of their research groups. I wish to make it clear that I am an entirely disinterested party who has received no remuneration of any kind for the production of this report.

Daniel Lunn
Department of Statistics
University of Oxford

Brief Curriculum Vitae and statement of qualifications

Date of Birth: 18 February 1942

Qualifications: M.A. (Oxon), D.Phil. (Oxon)

Professional bodies:
Fellow of the Royal Statistical Society
Member of the American Statistical Association
Member of the International Assocation for Statistical Education

Appointments:
1972 Department of Statistics, Faculty of Mathematics, The Open University
1985 Worcester College, Oxford (Fellow and Tutor in Mathematics).
2009 Emeritus Fellow, Worcester College, Oxford

Research areas of publications:
Multilevel modelling
Bayesian statistics (reliability, simulation, meta-analysis)
Medical statistics, clinical trials (categorical data, repeated measures, generalised estimating equations, generalised linear models, robust methods), meta-analysis.
Directional data (estimation in circular and spherical geometries, robust methods)
Analysis of spectroscopic data (discrimination, multivariate methods of data reduction)
Reliability (reliability of safety critical and other software)
Distribution theory (smoothing, simulation)
Graphical methods and displays.

Books:
Lunn, A.D. and McNeil, D.R. (1991): *Computer-Interactive Data Analysis*, Chichester: John Wiley.

Hand, D.J., Daly, F., Lunn, A.D.,McConway, K.J. and Ostrowski, E. (1994): *A Handbook of Small Data Sets*, London: Chapman and Hall.

Daly, F., Hand, D.J., Jones, M.C.,Lunn, A.D. and McConway, K.J. (1995): *Elements of Statistics*, London: Addison-Wesley.