A STATISTICAL ANALYSIS OF *THE PHILOSOPHICAL GOURMET REPORT* 2009

DAN HICKS

This note summarizes the methods and conclusions of my statistical analysis of *The philosophical gourmet report* 2009 (henceforth PHG).

1. Some prior critiques

My analysis was motivated by several prior critiques of PHG. The two most influential on me have been Julie Van Camp's 'Female-friendly departments: A modest proposal for picking graduate programs in philosophy' [1] and Richard's Heck's 'About PGR' [2].

One of Van Camp's primary concerns is that PHG exhibits gender bias:

> If there was a gender bias in judging the work of female researchers, then we might expect that departments which have a higher percentage of women on tenured/tenure-track appointments suffer in the rankings overall. Departments with higher percentages of women might be lower on the list than they should otherwise be, and departments with lower percentages of women might be higher on the list than they should otherwise be. If we saw a clear and consistent correlation (say, the lower the percentage of women on a faculty, the higher the ranking of the department on the list), that might raise reasonable suspicions about gender bias, but certainly not settle the matter. It would be even more suspicious if a department moved up or down on the list over the years in concert with increases and decreases in its proportion of faculty women.

While Van Camp does acknowledge that 'there is no obvious correlation', she still argues that there are still 'some data which are troubling on an impressionistic level' and calls for a more thorough study of the evaluators to look for gender bias. [3]

------

[1] Retrieved from `http://www.csulb.edu/~jvancamp/Female_Friendly.html` on Wednesday, May 6, 2009.

[2] Retrieved from `http://frege.brown.edu/heck/philosophy/aboutpgr.php` on Wednesday, May 6, 2009.

[3] Another of Van Camp's primary criticisms is that women are underrepresented on the PHG Advisory Board, even more than they are underrepresented in the discipline as a whole. Where approximately 20% of tenured and tenure-track faculty in philosophy are women, only 14.3% of the members of the Advisory Board of PHG 2009 are women.

Heck's concerns are more classically methodological: it is not clear that the survey methodology is actually measuring anything, much less anything that is an important discriminant between 'graduate programs in philosophy'.

My aim in this analysis was to determine, first, whether the PhG rankings were correlated in a statistically significant way with the percentage of women faculty, and second, what (other) factors were correlated in a statistically significant way with PhG rankings. Note that this means I am looking only at the outcomes of the ranking method; my analysis completely brackets the issues dealing with the ranking method itself.

2. The dataset

2.1. **Raw data.** My full dataset included all 98 departments listed in Van Camp's table of faculty women.[4] The initial or raw data included the PhG 2006 ranking, PhG 2009 ranking, and the percent of faculty women for each department. Because many departments on Van Camp's table were not ranked by PhG, and rankings stop at 54 in 2006, I assigned these departments a ranking of 55. I will refer to departments that were included in the PhG 2009 as 'ranked', and all others from Van Camp's table as 'unranked'.

The other variables used to build my dataset were PhG 2009 specialization rankings – for example, the ranking of departments in metaphysics. Specialization rankings are by 'groups', with group 1 being with highest and group 5 being the lowest. Since only a subset of ranked departments are ranked in any given specialization, I assigned all departments not ranked in that specialization (including those not ranked at all) to a group 6. I included all specializations ranking at least 37 departments, and a few 'more specialized' areas of my own personal interest. The specializations, and their variable names, are listed in table 1.

Specialization	Variable	Specialization	Variable
Metaphysics	Meta	Epistemology	Epist
Philosophy of mind	Mind	Philosophy of language	Lang
Ethics	Ethics	Metaethics	Methics
Applied ethics	Aethics	Political philosophy	Political
Philosophical logic	PhLogic	Mathematical logic	MLogic
Philosophy of math	Math	Philosophy of science	PhS
Philosophy of physics	PhPhys	Philosophy of biology	PhBio
Ancient	Ancient	17th Century	EM17
18th Century	EM18	Kant	Kant
Nineteenth Century	Nine	20th Century Continental	Cont
History of Analytic	HA	Feminist Philosophy	Fem

Table 1: Specializations and variables

2.2. **Derived variables.** Because of the way PhG rankings are calculated, a given department can move up or down in rankings purely as a result of changes in other departments. To control for this, I divided the rankings into 'decades'. The top 10 departments (ranked 1-10), for example, are decade 1, while the departments ranked 21-30 are decade 3. The resulting variable, PhG2009dec, was the primary dependent variable of my investigation.

For my analysis, I used a standard statistical technique, called *linear regression* or *ordinary-least-squares*. This technique takes one dependent (or, in philosophical jargon, explanandum) variable and a number of independent (or, we might say, explanans) variables, and calculates an optimal linear function relating all these. For example, suppose we have N observations $(\hat{z}_i, \hat{x}_i, \hat{y}_i), i = 1, \dots, N$, with dependent variable $\hat{Z} = (\hat{z}_i)^T$ and independent variables $\hat{X} = (\hat{x}_i)^T$ and $\hat{Y} = (\hat{y}_i)^T$ (by convention, the 'hat' is used to distinguish observed or measured values from true or actual values). A linear regression would return a function of the form

$$\hat{Z} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X} + \hat{\beta}_2 \hat{Y} + (e_i)^T,$$

where the β_j are called the *regression coefficients* and the e_i are error terms or *residuals*. The linear regression technique determines this function by minimizing the sum of the squares of the residuals,

$$SSE = \sum_{i=1}^{N} e_i^2,$$

which is the square of the magnitude of the error vector $(e_i)^T$.

This approach assumes that the independent variables are statistically independent – that is, that there is no significant linear relation between any two independent variables. For example, suppose that, corresponding to the example above, the true relation between the variables Z, X, Y is given by

$$Z = \beta_0 + \beta_1 X + \beta_2 Y + (\varepsilon_i)^T,$$

but also that

$$Y = \alpha X + (\gamma_i)^T,$$

where $\alpha \neq 0$ and $(\gamma_i)^T$ is statistically independent of X. Then Y can be 'reduced' to X in the equation for Z:

$$\begin{aligned} Z &= \beta_0 + \beta_1 X + \beta_2 Y + (\varepsilon_i)^T \\ &= \beta_0 + \beta_1 X + \beta_2(\alpha X + (\gamma_i)^T) + (\varepsilon_i)^T \\ &= \beta_0 + (\beta_1 + \beta_2\alpha)X + \beta_2(\gamma_i)^T + (\varepsilon_i)^T \end{aligned}$$

This means that the coefficients β_1, β_2 do not accurately represent the contributions of the independent variables to Z – some of the contribution of Y is actually a contribution of X.

The problem can be nicely illustrated by contrasting the results below with the preliminary results I announced on Facebook. In those results, I reported that most of the specialization rankings had dropped out of the regression, and the best model included only three independent variables: Metaphysics, Mind, and PhG 2006 decade. I later discovered that PhG 2006 decade was significantly correlated with a number of other specializations; hence the contribution of PhG 2006 decade

in the preliminary results actually reflected the contribution of a number of other specializations.

Fortunately, the problem is fairly easy to deal with. We define a new variable

$$Y' = Y - \alpha X = (\gamma_i)^T,$$

which is then independent of X, by construction.[5] Call Y' the *reduction* of Y by X, and say that Y is *reduced as* Y'. (This terminology is mine; I don't know of a standard term for this in statistics.)

Beyond the correlations with PHG 2006 decade mentioned above, it is *prima facie* plausible that many of the specialization rankings are not statistically independent. One can imagine, for example, that departments ranked highly in the philosophy of physics would, for that reason, also be ranked highly in the philosophy of science more generally. Hence, the actual specialization rankings were stored in variables of the form X_pool, where X is the variable as given in table 1. I then conducted regressions of smaller or more narrow specializations against more prominent or general specializations in the same area, as determined by the classification in PHG 2009. For example, Philosophy of mind was regressed against Metaphysics and Epistemology. The four areas are 'Metaphysics and epistemology', 'Philosophy of the sciences and mathematics', 'Theory of value', and 'History of philosophy'. Certain 'cross-cutting' specializations were regressed against a wider selection of other specializations. For example, Feminist philosophy was found to be correlated in a statistically significant way with Metaphysics and 17th century. These regressions were then used to define reduced specialization variables; it is these reduced variables that were used in all the regressions below. Finally, PHG 2006 decade was regressed against all the specialization variables, and reduced in the same way; the resulting variable was stored as r2006.

In the regressions, five test statistics were examined: R^2, p-value, and the Akike, Schwarz, and Hannan-Quinn criteria. R^2 is standardly interpreted as a measure of how much of the variance of an dependent variable can be attributed to variance of the independent variables; hence $R^2 = .95$ means that 95% of the variance in the dependent variable can be attributed to variance in the independent variables. The p-value of a value of a quantity is a measurement of how likely that value is to occur given a null hypothesis (often that the quantity is actually equal to 0). Hence a lower p-value indicates that the null hypothesis is more likely to be false. Conventionally, $p < 0.05$ is considered the threshold for statistical significance, and hence the threshold for rejecting the null hypothesis.

Calculations of p-values assume that the residuals fall in a normal distribution. If this is not the case, then a p-value calculation is not a reliable way to measure the quality of a calculation of the value of the quantity, while the last three test statistics can be used to compare two models for quality of fit. In most cases, residuals of regressions involving the full dataset were not normally distributed. I therefore conducted all of the regressions below with the subpopulation of ranked departments.

---

[5]The resulting process for constructing a set of statistically independent variables is basically the same as the process for constructing a minimal basis in linear algebra.

3. Gender bias

I first tested for gender bias by regressing the PhG 2009 decades against the percent faculty women. A normal distribution of the residuals required limiting my attention to the 53 ranked departments. This regression returned the results in model 1. They indicate a small, statistically insignificant correlation between

Model 1: OLS estimates using the 53 observations 1–53
Dependent variable: PhG2009dec

	Coefficient	Std. Error	t-ratio	p-value
const	2.66199	0.588752	4.5214	0.0000
PercWomen	0.0174652	0.0270020	0.6468	0.5207

Mean dependent var	3.018868	S.D. dependent var	1.487002
Sum squared resid	114.0456	S.E. of regression	1.495389
R^2	0.008136	Adjusted R^2	−0.011312
$F(1, 51)$	0.418364	P-value(F)	0.520655
Log-likelihood	−95.51086	Akaike criterion	195.0217
Schwarz criterion	198.9623	Hannan–Quinn	196.5371

Model 1: Test for gender bias

PhG 2009 decade and percent faculty women that accounts for almost none of the variance in PhG 2009 decade. Hence, it is reasonable to conclude that the data show no gender bias in the outcomes of the survey methodology: departments with a larger number of women are not thereby penalized.

4. Specialization and past ranking

Since gender bias does not appear to account for PhG ranking, I turned my attention to the other independent variables. Regressing against all the specialization rankings in my dataset and PhG 2006 decade returned the results in model 2. In this initial regression, most of the coefficients are small and not statistically significant. After testing the effects of removing and adding back in different combinations of variables, I found that the best overall model is model 3.

In particular, this model is substantially better than the overall best model with just specialization rankings, model 4. Since r2006 is, by construction, statistically independent of the specialization variables, this strongly suggests that PhG 2006 ranking makes an important independent contribution to PhG 2009 ranking.

The relative magnitude of the contribution of each independent variable can be determined by comparing their correlation coefficients in model 3, with the caveat that the differences between coefficients for most pairs of specialization variables is not statistically significant. However, the difference between the coefficient for r2006 and any of the specialization variables is statistically significant. That is, it is rather likely that the order in table 2 is not quite right, but the contribution of PhG 2006 rank is still much greater than the contribution of any of the specializations. Table 2 summarizes the contribution of each independent variable as a percentage of the overall variance in PhG2006dec. As this table indicates, the contribution of PhG

Model 2: OLS estimates using the 53 observations 1–53
Dependent variable: PhG2009dec

	Coefficient	Std. Error	t-ratio	p-value
const	−2.98813	1.03141	−2.8971	0.0071
Meta	0.241866	0.0734902	3.2911	0.0026
Epist	0.189740	0.0897620	2.1138	0.0433
Mind	0.298954	0.0687656	4.3474	0.0002
Lang	0.140997	0.0970280	1.4532	0.1569
Ethics	0.104283	0.0991664	1.0516	0.3017
Methics	−0.0185187	0.0908063	−0.2039	0.8398
Aethics	0.0841096	0.0821064	1.0244	0.3141
Political	−0.0714102	0.0790966	−0.9028	0.3741
PhLogic	0.276008	0.0879033	3.1399	0.0039
MLogic	0.107114	0.103542	1.0345	0.3095
Math	0.105122	0.143800	0.7310	0.4706
PhS	0.108632	0.0542031	2.0042	0.0545
PhPhys	0.128056	0.0690311	1.8551	0.0738
PhBio	0.0106231	0.0582315	0.1824	0.8565
Ancient	0.285972	0.0884595	3.2328	0.0031
EM17	0.248147	0.0566607	4.3795	0.0001
EM18	0.222010	0.0796084	2.7888	0.0092
Kant	0.0932682	0.0491631	1.8971	0.0678
Nine	−0.0789909	0.0693418	−1.1392	0.2640
Cont	−0.00936011	0.0775011	−0.1208	0.9047
HA	0.208849	0.0642122	3.2525	0.0029
Fem	0.210873	0.0615433	3.4264	0.0018
r2006	0.527140	0.105326	5.0048	0.0000

Mean dependent var	3.018868	S.D. dependent var	1.487002
Sum squared resid	6.003022	S.E. of regression	0.454973
R^2	0.947791	Adjusted R^2	0.906384
$F(23, 29)$	22.88967	P-value(F)	5.88e–13
Log-likelihood	−17.48598	Akaike criterion	82.97196
Schwarz criterion	130.2590	Hannan–Quinn	101.1563

Model 2: Initial regression against specializations

2006 rank is enormous, even after controlling for the effects of every specialization –
more than twice the contribution of every specialization except Philosophical logic,
and then only barely.

5. INTERPRETATION

In a posting on Facebook, I announced as preliminary results that PʜG rank-
ing does not appear to be biased against women, but does appear to be biased in
two other ways. The first was that it was biased in favor of certain specializations –

Model 3: OLS estimates using the 53 observations 1–53
Dependent variable: PhG2009dec

	Coefficient	Std. Error	t-ratio	p-value
const	−2.50977	0.804350	−3.1202	0.0035
Meta	0.206612	0.0589284	3.5062	0.0012
Epist	0.152674	0.0772079	1.9774	0.0555
Mind	0.282288	0.0563766	5.0072	0.0000
Ethics	0.167085	0.0712076	2.3465	0.0244
PhLogic	0.312355	0.0661339	4.7231	0.0000
MLogic	0.127918	0.0735140	1.7401	0.0902
PhS	0.100749	0.0469412	2.1463	0.0385
PhPhys	0.124708	0.0621782	2.0057	0.0522
Ancient	0.271032	0.0703644	3.8518	0.0004
EM17	0.245680	0.0484268	5.0732	0.0000
EM18	0.180686	0.0640661	2.8203	0.0077
Kant	0.0885662	0.0426565	2.0763	0.0449
HA	0.166198	0.0482656	3.4434	0.0014
Fem	0.189467	0.0524610	3.6116	0.0009
r2006	0.554032	0.0916689	6.0438	0.0000

Mean dependent var	3.018868	S.D. dependent var	1.487002
Sum squared resid	6.728022	S.E. of regression	0.426425
R^2	0.941486	Adjusted R^2	0.917764
$F(15, 37)$	39.68839	P-value(F)	3.23e–18
Log-likelihood	−20.50746	Akaike criterion	73.01491
Schwarz criterion	104.5396	Hannan–Quinn	85.13778

Model 3: Best overall regression

namely, Metaphysics and Ethics – and against others. The second was the contribution of past ranking, which (in the spirit of Philip Kitcher) I attributed to 'unearned authority'.

On closer inspection, and after turning to reduced (and hence properly statistically independent) variables, the first claim of bias must be highly qualified. A wide array of specializations – including all the major areas, viz., Metaphysics, Epistemology, Ethics, Philosophy of science, Logic, and the most prominent historical specializations – contribute to PHG rankings, and the differences of their contributions are (mostly) not statistically significant. Two criticisms can still be made, however. First, Continental philosophy does not appear to make a significant contribution. This is partly due to the reduction process (approximately 50% of the variance in Cont was eliminated by reduction), but Continental philosophers may still legitimately object that their contribution, independent of their work as historians, should be non-trivial. Second, certain relatively narrow specializations make contributions, but not others. For example, there seems no good reason why Philosophy of mind should have a significant contribution, and not Philosophy of biology.

Model 4: OLS estimates using the 53 observations 1–53
Dependent variable: PhG2009dec

	Coefficient	Std. Error	t-ratio	p-value
const	-5.12691	0.694128	-7.3861	0.0000
Meta	0.328776	0.0807087	4.0736	0.0002
Epist	0.309424	0.102415	3.0213	0.0043
Mind	0.365001	0.0784735	4.6513	0.0000
Lang	0.211005	0.104165	2.0257	0.0493
Ethics	0.330414	0.0912101	3.6226	0.0008
PhLogic	0.178443	0.0987839	1.8064	0.0782
MLogic	0.283337	0.0951725	2.9771	0.0049
Ancient	0.234025	0.0957928	2.4430	0.0190
EM17	0.189042	0.0641680	2.9460	0.0053
HA	0.126060	0.0677379	1.8610	0.0699
Fem	0.157901	0.0700526	2.2540	0.0296

Mean dependent var	3.018868	S.D. dependent var	1.487002
Sum squared resid	14.59607	S.E. of regression	0.596659
R^2	0.873057	Adjusted R^2	0.838999
$F(11, 41)$	25.63446	P-value(F)	6.30e–15
Log-likelihood	-41.03095	Akaike criterion	106.0619
Schwarz criterion	129.7054	Hannan–Quinn	115.1540

Model 4: Best overall regression using only specialization rankings

Variable	Contribution	Variable	Contribution
r2006	16.7	PhLogic	9.4
Ancient	8.2	EM17	7.4
Mind	7.0	Meta	6.2
Fem	5.7	EM18	5.5
Ethics	5.0	HA	5.0
Epist	4.6	MLogic	3.9
PhPhys	3.8	PhS	3.0
Kant	2.7		
$total=R^2$	94.1		

Table 2: Contribution of statistically significant factors to PhG 2009 ranking, by percent variance in PhG2009dec

The second claim of bias, on the other hand, is essentially unchanged. If I am right in calling the contribution of prior ranking unearned authority, then this represents a contribution that is not based on actual academic achievement or quality. Since r2006 has been reduced by every specialization variable, it is certainly not clear what kind of actual academic achievement or quality it could represent. And this contribution is not just significant but significantly larger than the contribution of any variable that is at least nominally based on actual academic achievement or

quality. Nearly 17% of a department's PHG ranking appears to be based solely on its unearned authority.

As this analysis has looked solely at the results of the survey and ranking process, it cannot serve as the basis for making direct methodological recommendations. In particular, it cannot identify the causes of the problems I have identified here. It does, however, suggest that a methodological critique should look at the kind of contribution made by past ranking, including, first, determining whether or not my interpretation of this contribution as 'unearned authority' is accurate and, if so, second, how unearned authority plays such a large role in the survey methodology, and third, how this role can be diminished or eliminated.