STUDY “PHD”
STATISTICAL ANALYSIS

ARS STATISTICA SPRL Fils Jean-François
Boulevard des Archers, 40
1400 Nivelles
Tel: 0476/316.048
E-Mail: jean-francois.fils@ars-statistica.com
TVA: 534.861.760
Numéro d'entreprise: 0534.861.760
Web : www.ars-statistica.com

Dr Dupont
Service recherche
CHU Bruxelles
0, Rue Bidon
1000 Bruxelles
Tél.: 32 67 55 73 97
E-mail: Dr.Dupont@chu-bruxelles.be

Table of contents

1. Goal of the report 3
2. Descriptive statistics 3
2.1. Events 3
2.2. Log(Events) 4
2.3. Exposure 4
2.4. Log(Exposure) 5
2.5. Discrete variables 5
3. Research Questions 6
3.1. T-Test/ANOVA presentation of the data 6
3.2. T-Test example: exposure by residence group 7
3.3. ANOVA example: event by University group 7
3.4. ANOVA example: log(event) by University group 9
3.5. Logistic regression example: predict residence 10
3.6. Bibliography 10

1. Goal of the report

The goal of this report is to present the statistical results of the study PhD as an illustrative example, indicating the way Jean-François Fils works with medical doctors. The R version 2.15.2 was used to produce statistical results. The data presented in this report are available here http://data.princeton.edu/wws509/datasets/#phd. The Time to PhD data are available in a file containing five columns:

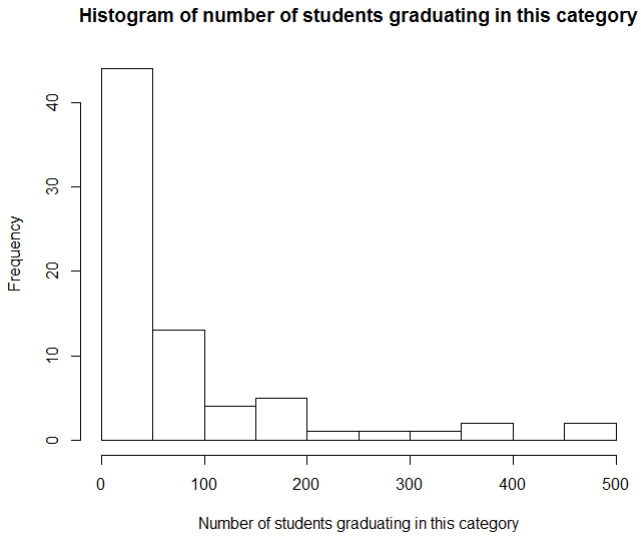
- year: coded 1 to 14, representing years of graduate school.
- university: coded Ber for Berkeley, Col for Columbia, Pri for Princeton.
- residence: coded 1 for permanent residents, 2 for temporary residents.
- events: number of students graduating in this category.
- exposure: number of person-years of exposure to graduation in this category.

Reference: Espenshade, T.J. and Rodríguez, G. (1997). Completing the Ph.D.: Comparative Performances of U.S. and Foreign Students. *Social Science Quarterly*, **78**:593-605.

2. Descriptive statistics

2.1. Events

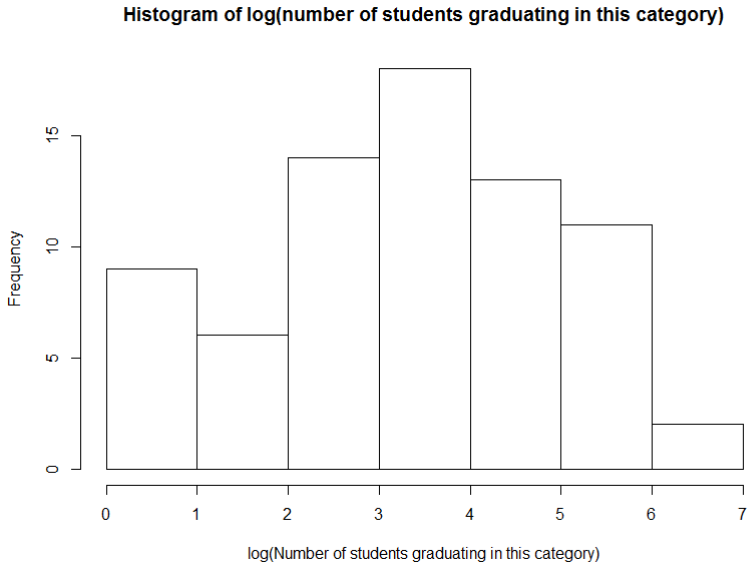
Nb Obs	Min.	1st Qu.	Median	mean	3rd Qu.	Max.	Var.	Std.
73	1.00	11.00	36.00	78.07	97.00	500.00	12383.51	111.2812



The mean event is 78.07, with a standard deviation equal to 111.2812.

2.2. Log(Events)

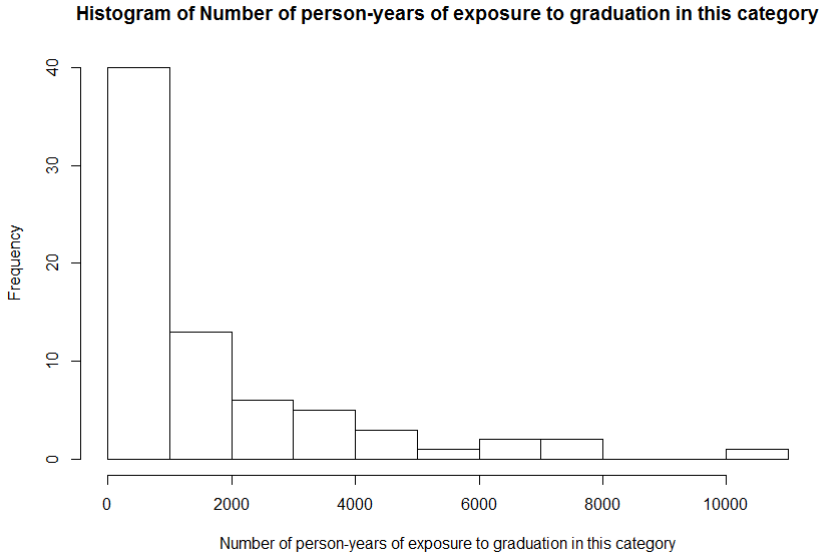
Nb Obs	Min.	1st Qu.	Median	mean	3rd Qu.	Max.	Var.	Std.
73	0.000	2.398	3.584	3.352	4.575	6.215	2.615238	1.61717



The mean log(event) is 3.352, with a standard deviation equal to 1.62.

2.3. Exposure

Nb Obs	Min.	1st Qu.	Median	mean	3rd Qu.	Max.	Var.	Std.
73	37	435	805	1750	2384	10600	4292469	2071.834



The mean exposure is 1750, with a standard deviation equal to 2071.834.

2.4. Log(Exposure)

Nb Obs	Min.	1st Qu.	Median	mean	3rd Qu.	Max.	Var.	Std.
73	3.611	6.075	6.691	6.850	7.777	9.268	1.38	1.1748

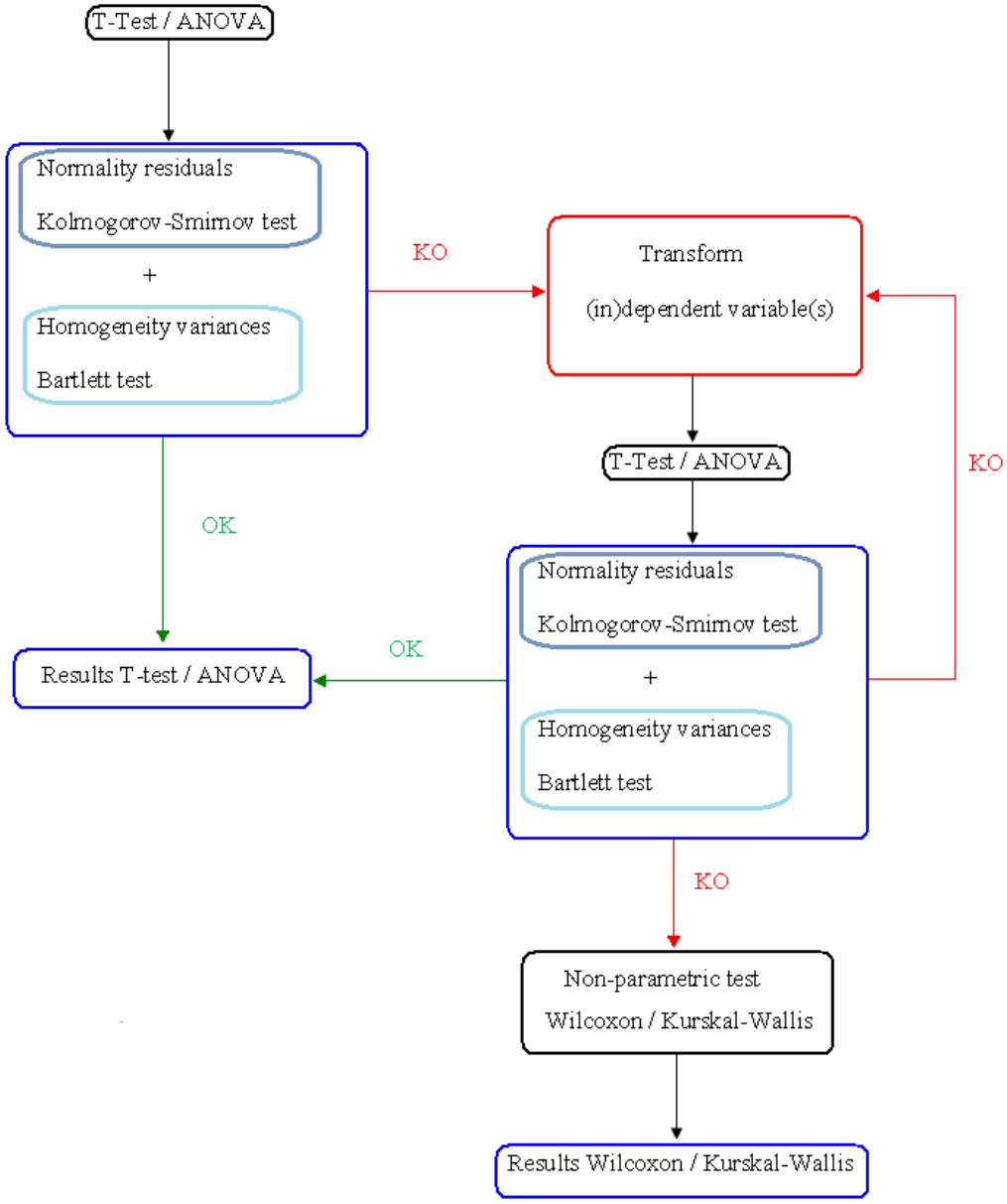
The mean log(exposure) is 6.850, with a standard deviation equal to 1.1748.

2.5. Discrete variables

Three Universities are present in the data: there are 28 observations for Berkley, 23 for Columbia and 22 for Princeton. There are 42 permanent residents and 31 temporary residents.

3. Research Questions

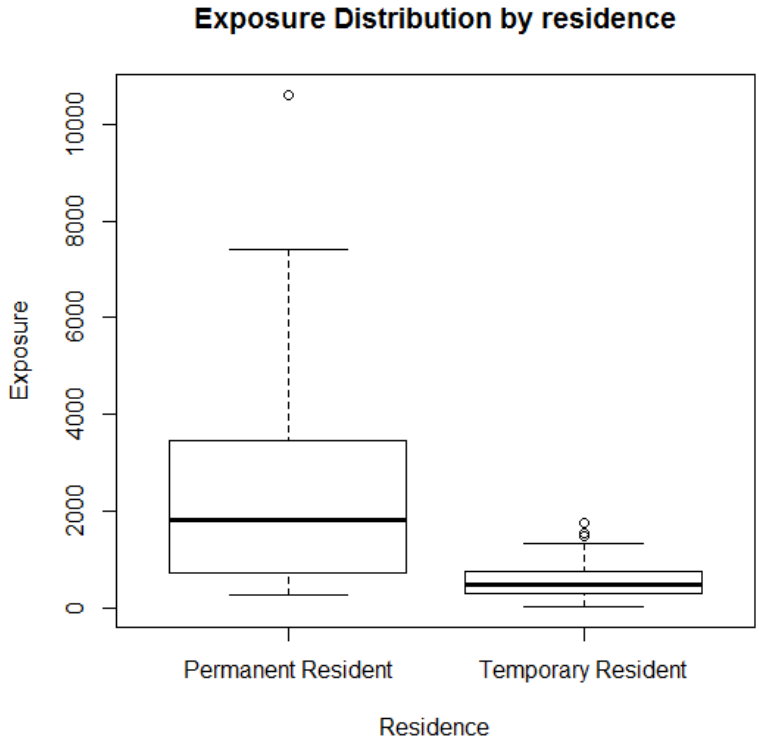
3.1. T-Test/ANOVA presentation of the data



The above picture presents the way the statistical tests (T-test and ANOVA) are presented hereunder. First, a T-test/ANOVA is performed to compare means between two or more groups. Next, assumptions of the T-test/ANOVA are tested: normality of the residuals and homogeneity of the variances. If the normality assumption is not met, we transform the dependent variable by taking, for example, the logarithm of it or its exponential. Indeed, a T-test/ANOVA can be performed for heterogeneous variances. Next, T-test/ANOVA is performed on the transformed data and the assumptions of the T-test/ANOVA are tested on this last model. If the assumptions are met, results of the T-test/ANOVA are presented. If they are not met, we have to perform the non-parametric equivalent of the T/test/ANOVA: Wilcoxon/Kurskal-Wallis test, respectively.

3.2. T-Test example: exposure by residence group

The Bartlett's K-squared (χ^2 =62.3016, df = 1, p-value < 0.001) indicates that the variances of exposure in the two groups are significantly different. The Kolmogorov-Smirnov (0.1753, p-value = 0.01967) indicates that the residuals of the model are not normally distributed. We tried several transformations of the dependent variable but our residuals were not normally distributed. We therefore use a non-parametric test: the Wilcoxon signed-rank test. Results (W = 1076, p-value < 0.001) indicate that there are significantly more person-years of exposure to graduation for permanent residents (median=2599.29) than for temporary residents' (median=599.42). The next boxplot depicts the results.



3.3. ANOVA example: event by University group

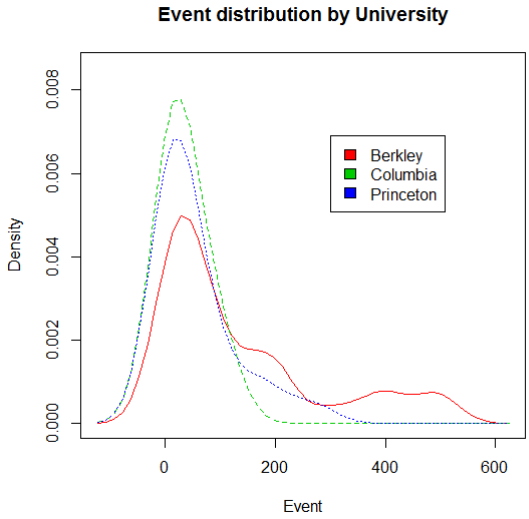
We performed an ANOVA with event as dependent variable and with University as independent variables. The results are presented hereunder.

	Df	Sum of Sq	Mean Sq	F value	Pr(F)
University	2	14464	7232	6.777	0.00204
Residuals	70	746971	10671		

Results indicate that University is significantly associated with the target (p-value=0.002). In other words, there are differences between Universities in terms of events.

	Diff	Lwr	upr	p adj
Columbia-Berkley	-100.04969	-169.65957	-30.439814	0.0027831
Princeton-Berkley	-79.57143	-150.04452	-9.098333	0.0231080
Princeton-Columbia	20.47826	-53.28833	94.244850	0.7846200

The Tuckey's multiple comparisons indicate that there are significantly less events in Columbia ($\bar{x} = 33.52$) than in Berkley ($\bar{x} = 133.57$, p-value = 0.002). There are also less events in Princeton ($\bar{x} = 54.00$) than in Berkley (p-value=0.023). Last, the comparison between Princeton and Columbia is not significant (p-value=0.78). The next picture depicts the results:



The Bartlett's K-squared (χ^2 =43.1254, df = 2, p-value < 0.001) indicates that the variances of event in the three groups are significantly different. The Kolmogorov-Smirnov test indicates that the residuals are normally distributed (D = 0.2055, p-value = 0.09171).

Because the variances are not equal between groups, we perform the same ANOVA with the log(event) and we will check if our homoscedasticity assumption holds.

3.4. ANOVA example: log(event) by University group

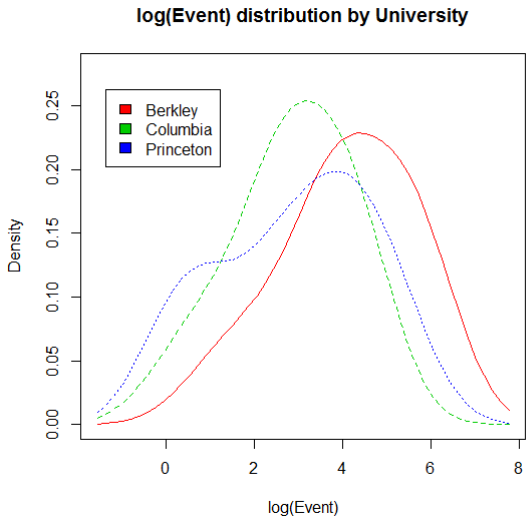
We performed an ANOVA with log(event) as dependent variable and with University as independent variables. The results are presented hereunder.

	Df	Sum of Sq	Mean Sq	F value	Pr(F)
University	2	24.89	12.446	5.332	0.007
Residuals	70	163.40	2.334		

Results indicate that University is significantly associated with the target (p-value=0.007). In other words, there are differences between Universities in terms of log(events).

	Diff	Lwr	Upr	p adj
Columbia-Berkley	-1.22132740	-2.250885	-0.1917694	0.0160476
Princeton-Berkley	-1.17856572	-2.220891	-0.1362403	0.0228769
Princeton-Columbia	0.04276169	-1.048276	1.1337992	0.9951560

The Tuckey's multiple comparisons indicate that there are significantly less log(events) in Columbia ($\bar{x} = 2.87$) than in Berkley ($\bar{x} = 4.09$, p-value = 0.016). There are also less log(events) in Princeton ($\bar{x} = 2.91$) than in Berkley (p-value=0.023). Last, the comparison between Princeton and Columbia is not significant (p-value=0.99). The next picture depicts the results:



The Bartlett's K-squared (χ^2 =1.6364, df = 2, p-value = 0.4412) indicates that the variances of log(events) in the three groups are not significantly different. The Kolmogorov-Smirnov test indicates that the residuals are normally distributed (D = 0.0959, p-value = 0.8904).

3.5. Logistic regression example: predict residence

We tried to predict residence with three independent variables: event, exposure and university. We used a backward selection strategy, meaning that we removed one by one non-significant variables of the model. The only variable associated with residence is exposure.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.509755	0.490504	3.078	0.002
Exposure	-0.001651	0.000492	-3.355	<0.001

The estimate is negative, indicating that less exposure was associated with a temporary residence.

3.6. Bibliography

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.