

Epidemiologic surveillance and impact evaluation¹

Sorveglianza Epidemiologica e Valutazione di Impatto

Annibale Biggeri ²

Dipartimento di Statistica “G. Parenti”, Università di Firenze

Unità di Biostatistica, CSPO, Firenze

e-mail: abiggeri@ds.unifi.it

Dolores Catelan

Dipartimento di Statistica “G. Parenti”, Università di Firenze

e-mail: catelan@ds.unifi.it

Emanuela Dreassi

Dipartimento di Statistica “G. Parenti”, Università di Firenze

e-mail: dreassi@ds.unifi.it

Riassunto: La sorveglianza epidemiologica è definita come monitoraggio e individuazione di tassi di mortalità alterati per una certa popolazione a rischio. Consiste nell’analizzare l’andamento della frequenza di eventi di interesse sanitario come i tassi di mortalità nel tempo e nello spazio. Gli approcci statistici adottati per la sorveglianza epidemiologica seguono sostanzialmente due approcci: quello legato al controllo della qualità (ad esempio tecniche CUSUM) e quello diretto all’identificazione di anomalie nei dati (ad esempio test di clustering). Il presente lavoro propone un modello statistico basato sul “positive false discovery rate” per identificare l’eccesso di rischio nel contesto di investigazioni epidemiologiche di tipo descrittivo.

Keywords: Bayesian Mixture Models, False Discovery Rate, Large Tables, q-value

1. Introduction

Epidemiologic surveillance is defined as monitoring and detection of altered disease rates within a given population at risk (see McMahon and Pugh, 1970). Measuring occurrence variation by space, time or space-time is a specific task which may require to set up a disease registry (dos Santos Silva, 1999 chp 17).

The statistical approaches to surveillance follow two main streams: quality control (e.g. CUSUM techniques, see Frisén and De Marè, 1991), detection of extremes (e.g. test for clustering, Kulldorff *et al.*, 2007).

Natural history of the disease and exposure distribution is usually known and a flow chart of public health actions for each possible results of the surveillance activity is assumed (e.g. the national mesothelioma registers; for Italy Decreto del Presidente del Consiglio dei Ministri n. 308 del 10 dicembre 2002, GU n. 31 del 7 febbraio 2003).

Stakeholders and decision makers however need measures of disease impact. Attributable cases and quality/disability adjusted years of life are used by WHO to assess global burden of disease (see Murray and Lopez, 1996).

¹COFIN-2006 (PRIN 2006131039).

²Address of correspondence: Viale Morgagni, 59, 50134 Firenze, Italia

Multiple testing in the framework of epidemiological studies or surveillance is rarely addressed. Several authors (see Rothman and Greenland, 1998) deny the adoption of any procedure to control family wise error rate. Within epidemiologic surveillance literature, global control of type I error is pursued (Frisén, 2003; Kulldorff, 2001; Elliott *et al.*, 2000).

Some authors (see Richardson *et al.*, 2004) discussed the low sensitivity of Hierarchical Bayesian modelling in disease mapping exercises, others (see Rolka *et al.*, 2007) noticed the cost in sensitivity of adopting a family wise error rate in monitoring. The last mentioned control of the False Discovery Rate (Benjamini and Hochberg, 1995).

We can do much better.

The paper aims to present a statistical model for identification of excesses in the context of descriptive epidemiological investigations.

The results are presented as attributable cases.

2. Motivating example

In 2005 an Epidemiological report has been requested by the Secretary of Health, Hygiene and Social Welfare of the Sardinia Region (Italy). It has been carried out by the Regional Epidemiological Observatory within the domain of ESA (Epidemiology Development and Environment) and with the support of the European Union (Biggeri *et al.*, 2006). The purpose of the report was to screen health status of resident populations in a priori identified areas at high environmental pressure. This required a common definition of health indicators among areas and standardized statistical approaches to data analysis. A total of 73 municipalities accounting for a population of 917.977 in 2001 census (56% of the total population of Sardinia) were considered. Mortality data (1981-2001) and hospital discharge data (2001-2003) were collected. Three steps of analysis were performed: Descriptive analysis for the most recent time period available (11 indicators by 35 causes of death and 48 hospital discharge diagnoses, gender specific); Geographical analysis (parametric Homogeneity test, Bayesian posterior means and posterior probabilities); Temporal analysis of mortality data for the whole period 1981-2001 and Birth cohort analysis, 1904-1948, with estimates of cumulative risks and cumulative SMRs in age 30-74.

The choice of appropriate reference population were evaluated comparing regional, material deprivation, and “local” adjusted indicators (Sokal and Rohlf, 1981, Bland and Altman, 1999). Localized excess or deficit are more likely attributable to environmental exposures: we used negative binomial Homogeneity test (Saha and Paul, 2005) and posterior probabilities (Richardson *et al.*, 2004). Age-period and age-cohort analysis were done (Clayton and Schifflers, 1987) looking for absolute and relative trends. Cohort effects were quantified by cumulative risks and cumulative SMRs (Breslow and Day, 1987; Coleman *et al.* 1993). Confidence bands were obtained by parametric bootstrap (Efron and Tibshirani, 1993).

A large number of comparisons were made. Overall we analysed 84 disease codes - 36 for mortality and 48 for hospital admissions - by 18 areas resulting in 1512 Relative Risk estimates.

How do address such huge number of multiple comparisons ?

3. Methods

3.1 Multiple inference

Generally speaking a specific procedure when performing multiple comparisons is advocated to proper control the probability of a false positive result. When doing a single hypothesis testing we maximize power conditional on type I error rate (PCER approach: Per Comparison Error Rate). While when multiple hypothesis testing is performed we aim to control the overall error rate at a desired level while maintaining the power of each test as much as possible. The commonly controlled quantity is FWER: Family Wise Error Rate. FWER is the probability of yielding one or more false positives out of all the hypothesis tested (Bonferroni, 1936).

There are situations in which major interest is in the rate of false positive among all rejected hypotheses rather than the probability of making one or more type I errors (FDR, Benjamini and Hochberg, 1995).

Indeed, the authors comment that control of FWER is important when our decision based on the whole set of individual inferences is likely to be erroneous when at least one of them is. This may be the case when several areas/populations are compared against a reference population (i.e. regional average), and a single area is considered out of the set of areas which are declared significantly better/worse than the standard. However, a given area/population and a reference are often compared by testing relative risks for various diseases (different end points in clinical trials terminology). The overall conclusion that the area is at higher/lower risk need not be erroneous even if some of the null hypotheses are falsely rejected.

Considering the possible outcomes from m hypothesis tests (see Table 1) we can define the followings.

The family wise error rate (FWER) (Bonferroni):

$$P(V \geq 1)$$

The False Discovery Rate (FDR) (see Benjamini and Hochberg, 1995):

$$E(V/R \cup 1)$$

However, FDR can be expressed also as:

$$E\left(\frac{V}{R \cup 1}\right) = P(R > 0) E\left(\frac{V}{R} \mid R > 0\right)$$

which provides a link to the positive False Discovery Rate (pFDR), defined as (Storey, 2003)

$$E\left(\frac{V}{R} \mid R > 0\right)$$

pFDR has a Bayesian interpretation and has a connection to classification theory.

As noted by Storey (2002) in the FDR procedure the error rate is controlled for all values of the number of true null hypotheses simultaneously without using any information in the data about it. But we have information on $P(R > 0)$ and we use it when considering positive FDR (see also Storey and Tibshirani, 2001 for discussion of FDR vs pFDR).

pFDR is a posterior probability:

$$P(H_0 | T) = \frac{\pi_0 P(T | H_0)}{\pi_0 P(T | H_0) + \pi_1 P(T | H_1)}$$

(see Efron *et al.*, 2001, Tusher *et al.*, 2001 for an example in genomics data analysis).

Let us define a three level hierarchical Bayesian model on our example:

- level 1 $Y_{ik} \sim \text{Poisson}(\lambda_{ik})$; where area=1,... $i..I$, disease=1,... $k...K$ and $\log(\lambda_{ik}) = \theta_{ik} + \text{offset}_{ik}$, the offset being the log expected counts or person years at risk;
level 2 $\theta_{ik} = Z_{ik}\mu_{1ik} + (1 - Z_{ik})\mu_{2ik}$ where $\mu_{1ik} = 0$ and $\mu_{2ik} \sim \text{Normal}(\cdot, \tau_{ik})$
level 3 $Z_{ik} \sim \text{Bernoulli}(\pi_{ik})$ and $\tau_{ik} \sim \text{Gamma}(0.5, 0.005)$

Posterior classification probabilities are then used to thresholding significant diseases/areas, following Efron *et al.*, 2001.

However posterior probabilities from the bayesian Mixture model were denoted by the last authors "local FDR" :

$$\pi_{ik} | \theta_{ik} = P(H = 0 | T = \theta_{ik})$$

while, more consistently, Storey, 2003 define positive FDR as:

$$p\text{FDR}(\gamma) = P(H = 0 | T \in \Gamma_\gamma)$$

and by analogy with the p-value:

$$p\text{value} = P(T \geq \theta_{ik} | H = 0)$$

the q-value, directly related to the posterior probability of classification:

$$q\text{value} = P(H = 0 | T \geq \theta_{ik})$$

The estimation of pFDR and q-value is outlined in Storey, 2002. First, get an estimate of π_0 from the set of observed p-values:

$$\hat{\pi}_0(\lambda) = \frac{\#(p_i > \lambda)}{(1 - \lambda)m}$$

then,

$$p\hat{\text{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{[\#(p_i \leq \gamma)/m][1 - (1 - \gamma)^m]}$$

and,

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} p\text{FDR}(\Gamma_\alpha)$$

3.2 Large tables

A different situation arises when we consider a large collection of Relative Risks, under the expectation of few excesses.

Let assume:

$$O_{ik} \sim \text{Poisson}(E_{ik}\lambda_{ik})$$

where i indexed areas and k diseases. Then

$$\log(\lambda_{ik}) \sim \text{Normal}(\mu_k, \tau_k)$$

The posterior estimates of the relative risk (or log relative risk) have approximate Gaussian distribution and can be used to fill in the quantile-quantile plot (Carpenter *et al.*, 1997; Law *et al.*, 2001).

Law *et al.* (2001) suggest a visual approach to identify anomalous points. It consists in plotting the ordered $\log(\hat{\lambda}_{ik})$ against the expected order statistics. In the same plot we can draw "guide rails", that is the limits within which the plotted points would be expected to fall, if the plotted points would have been a sample from a Normal distribution.

Consider the largest of n draws from a $\text{Normal}(0, 1)$:

$$P(Z_n \leq z) = \Phi(z)^n$$

then

$$1 - \alpha = \Phi(z_\alpha)^n$$

$$z_\alpha = \Phi^{-1} \left[(1 - \alpha)^{1/n} \right]$$

Guides rails can be put at $z_\alpha \tilde{\sigma}$; where $\tilde{\sigma}$ is the standard deviation (estimated by the slope in the central 80% of the QQ-plot). See also Law *et al.*, 2001 for a bootstrap approach which avoid Gaussian approximations.

4. Results

Let us extract from the Sardinia report (Biggeri *et al.*, 2006 pages 41 and 71) the tables for two areas with some disease excesses: Portoscuso and La Maddalena. In table 2 and table 3 we report the p-value and q-value for Portoscuso and La Maddalena for the diseases that are statistically significant under a type I error of 10%. For Portoscuso, if we control False Discovery Rate at the level of 10% six diseases in the table would be declared significant. Notice that while the p value is the minimum false positive rate that is incurred calling each single test significant, the q-value measures the minimum false discovery rate for that test. That is, over all the tests that we declare significant, the percentage of them being false positive.

We could conclude that for Portoscuso out of 6 diseases declared as significant 1 of them is a false positive (yet we do not know which one). For La Maddalena, fixing a false discovery rate at 20% 4 diseases are declared significant, 1 of them being a false positive.

The overall proportion of the true null hypothesis π_0 is estimated (Storey, 2002). Figure 1 and Figure 2 report estimates for π_0 for the Portoscuso and La Maddalena examples, respectively. The purpose of the plot is also to gauge how reliable the estimate of π_0 is. Basically, a tuning parameter λ has to be chosen to estimate π_0 . A spline is added to the plot in order to facilitate the interpretation. The proportion of significant results in the study can be derived by $\pi_1 = 1 - \pi_0$.

For Portoscuso the estimated π_0 is 77% thus we have a 23% of altered Relative Risk. The estimate is stable over the range of λ . The situation is much more unstable for La Maddalena: we estimate a π_0 of 62% but this is clearly inappropriate.

Using the guide-rail approach we plot ordered posterior logRR against ordered standard Gaussian deviate. In the figures 3 and 4 asymptotic 80% guide-rails were added (Law *et al.* (2001)).

5. Discussion

In the literature the following approaches to Epidemiological Large Tables analysis are reported:

- graphical approach. QQ-plot of Empirical Bayes estimates and guide rails, asymptotics or bootstrapped (Carpenter *et al.*, 1997 and Law *et al.*, 2001);
- Bayesian multivariate inference. Posterior means and credibility intervals are derived and used to thresholding areas (Assunção and Castro, 1995);
- League tables. Multilevel or Bayesian approaches led to posterior ranks (Goldstein and Spiegelhalter, 1996);
- hierarchical Bayesian modeling. In particular triple goal estimates (Shen and Louis (1998)); and limited shrinkage estimators (Devine and Louis (1994)).

We included in our analysis the simple graphical approach, listed before. However it is clearly appropriate to screen for few extremes rates or risks and can fail to detect a situation with a large number of altered disease rates or, for a given disease, areas. The other approaches are variation within a general hierarchical Bayesian modeling framework.

We considered two hierarchical Bayesian models fitted to the large table of relative risks. Coherently, we adopted a Bayesian approach to multiple inference (Storey, 2003) and found pFDR consistent with posterior ranking.

The first model specified was a three level mixture model, where the interest was in posterior classification probabilities. Spatial random terms could have been specified when considering areas on a continuous region, while for correlated responses see Assunção and Castro (1995). Both were inappropriate in our example.

q-values can be calculated using the procedure of Storey (2002). This approach is independent on the model used to get p-values. In the application here we considered Poisson maximum likelihood estimates for each area and each disease in turn.

The results are consistent. When looking at Bayesian posterior classification probabilities we are more conservative. The model fitted to the overall table showed some shrinkage effect, as expected. The interpretation is straightforward: if the interest lies in each single comparison, the procedure based on Poisson likelihood should be preferred. When exploring the set of compromised diseases by area, the Bayesian approach has the minimum loss.

Looking at the guide-rails approach, the Bayesian model specified was a two level hierarchical Poisson-Gaussian. If needed, random terms, spatially structured, could have been

specified. Posterior log RR are then graphically displayed and QQ-plot obtained.

The procedure is easy and nice to report. However when it is appropriate? Suppose in a given area the environmental pressure resulted in all disease rates above the reference population rates. Then the QQ-plot will fail to detect any discrepancy. Only changes in shape of the RR distribution will be identified.

The results presented supported this view: Portoscuso is a high affected area with several disease with risk above the reference. The guide-rails approach identifies only the extreme one (pneumoconiosis), in contrast with q-values estimates of 23% altered RRs.

6. Conclusions

Spatial epidemiological surveillance is based on disease mapping, cluster detection and focused analyses. Family Wise Error Rate control is usually advocated, and a discussion of alternatives to improve sensitivity is in Rolka *et al.* (2007)) (pages 1838-1839, and 1852). Our motivating example is in Biggeri *et al.* (2006). We showed how to better screen large tables of relative risks for altered disease occurrence. In the example 84 disease codes by 18 areas resulting in 1512 Relative Risk estimates were examined. A Bayesian approach provides useful to identify the set of anomalous associations. We considered in particular Bayesian equivalent of p-values (q-values, Storey, 2003) to threshold areas or diseases and compare to posterior statistics or guide-rails (Law *et al.*, 2001) approaches.

References

- Assunção R. and Castro M. (1995) Multiple cancer sites incidence rates estimation using a multivariate bayesian model, *International Journal of Epidemiology*, 33, 508–516.
- Benjamini Y. and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B*, 57, 289–300.
- Biggeri A., Lagazio C., Catelan D., Pirastu R., Casson F. and B. T. (2006) Environment and health in high risk areas of sardinia, italy, *Epidemiologia e prevenzione*, 30, S1–S96.
- Bland J. and Altman D. (1999) Measuring agreement in method comparison studies, *Statistical Methods in Medical Research*, 8, 135–160.
- Bonferroni C. (1936) Teoria statistica delle classi e calcolo delle probabilit, *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.
- Breslow N. and Day N. (1987) *Statistical Methods in Cancer Research. Vol. II, The Design and Analysis of Cohort Studies*, International Agency for Research on Cancer, Lyon.
- Carpenter L., Maconochie N., Roman E. and Cox D. (1997) Examining associations between occupation and health by using routinely collected data, *Journal of the Royal Statistical Society A*, 160, 507–521.
- Clayton D. and Schifflers E. (1987) Models for temporal variations in cancer rates. iii, *Statistics in Medicine*, 6, 449–481.
- Devine O. and Louis T. (1994) A constrained empirical bayes estimator for incidence rates in areas with small populations, *Statistics in Medicine*, 13, 1119–1133.
- dos Santos Silva I. (1999) *Cancer Epidemiology. Principles and Methods*, IARC, Lyon.

- Efron B. and Tibshirani R. (1993) *An introduction to the bootstrap*, Chapman and Hall.
- Efron B., Tibshirani R., Storey J. and Tusher V. (2001) Empirical bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, 96, 1151–1160.
- Elliott P., Wakefield J., Best N. and Briggs D. (2000) *Spatial Epidemiology: Methods and Applications*, Oxford.
- Frisén M. (2003) Statistical surveillance, optimality and methods, *International Statistical Review*, 71, 403–434.
- Frisén M. and De Marè J. (1991) Optimal surveillance, *Biometrika*, 78, 271–280.
- Goldstein H. and Spiegelhalter D. (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society A*, 159, 385–443.
- Kulldorff M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society A*, 164, 61–72.
- Kulldorff M., Mostashari F., Duczmal L., Yih W., Kleinman K. and Richard Platt R. (2007) Multivariate scan statistics for disease surveillance, *Statistics in Medicine*, 26, 1824–1833.
- Law G., Cox D., Machonochie N., Simpson J., Roman E. and Carpenter L. (2001) Large tables, *Biostatistics*, 2, 163–171.
- McMahon B. and Pugh P. (1970) *Epidemiology: principles and methods*, Boston Little Brown and Co.
- Murray C. and Lopez A. (1996) *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020*, Harvard University Press, WHO, The World Bank.
- Richardson S., Thomson A., Best N. and Elliott P. (2004) Interpreting posterior relative risk estimates in disease-mapping studies, *Environmental Health Perspective*, 112, 1016–1025.
- Rolka H., Burkom H., Cooper G., Kulldorff M., Madigan D. and Wong W.K. (2007) Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs, *Statistics in Medicine*, 26, 1834–1856.
- Rothman K. and Greenland S. (1998) *Modern Epidemiology*, Lippincott Williams and Wilkins.
- Saha K. and Paul S. (2005) Bias corrected maximum likelihood estimator of the negative binomial dispersion parameter, *Biometrics*, 61, 180–186.
- Shen W. and Louis T. (1998) Triple-goal estimates in two-stage hierarchical models, *Journal of the Royal Statistical Society B*, 60, 455–471.
- Sokal R. and Rohlf F. (1981) *Biometry. Second edition*, Freeman, New York.
- Storey J. (2002) A direct approach to false discovery rates, *Journal of the Royal Statistical Society B*, 64, 479–498.
- Storey J. (2003) The positive false discovery rate: a bayesian interpretation and the q-value, *The Annals of Statistics*, 31, 2013–2035.
- Storey J. and Tibshirani R. (2001) *Estimating false discovery rates under dependence, with applications to DNA microarrays*, Technical Report, Stanford University.
- Tusher V., Tibshirani R. and Chu G. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Sciences*, 98, 5116–5121.

Table 1: Possible outcomes from m hypothesis tests

	Accept null	Reject null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Table 2: The q -values for Portoscuso, males 1997-2001. In bold $FDR \leq 0.10$ q -values are reported.

disease	p value	q value
Pneumoconiosis	0	0
Diabetes	0.00166	0.0151
Prostate cancer	0.00202	0.0151
Lung cancer	0.01180	0.0661
Respiratory acute	0.01505	0.0675
CHD	0.01848	0.0690
Liver cirrhosis	0.03451	0.1098
Ill defined conditions	0.03918	0.1098
Respiratory chronic	0.04689	0.1116
Melanoma	0.04980	0.1116
Larynx cancer	0.06508	0.326

Table 3: The q -values for La Maddalena, males 1997-2001. In bold $FDR \leq 0.20$ q -values are reported.

disease	p value	q value
CHD	0.00653	0.0981
Pleural Tumor	0.01090	0.0981
Respiratory chronic	0.02151	0.1039
non Hodgkin Lymphoma	0.02310	0.1039
Bladder cancer	0.06960	0.2163
Infectious diseases	0.07210	0.2163

Figure 1: Estimates of π_0 for Portoscuso example

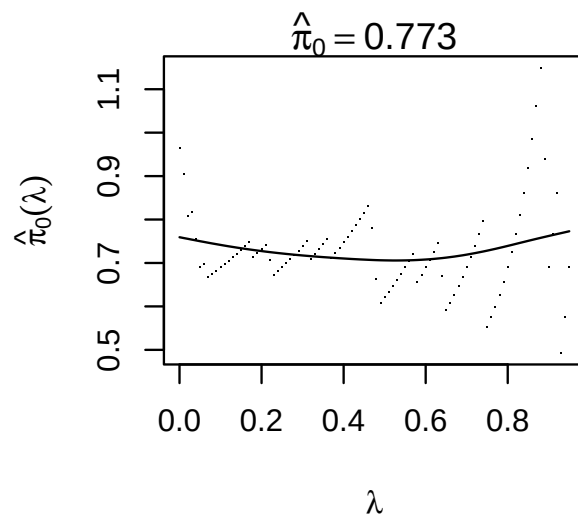


Figure 2: Estimates of π_0 for La Maddalena example

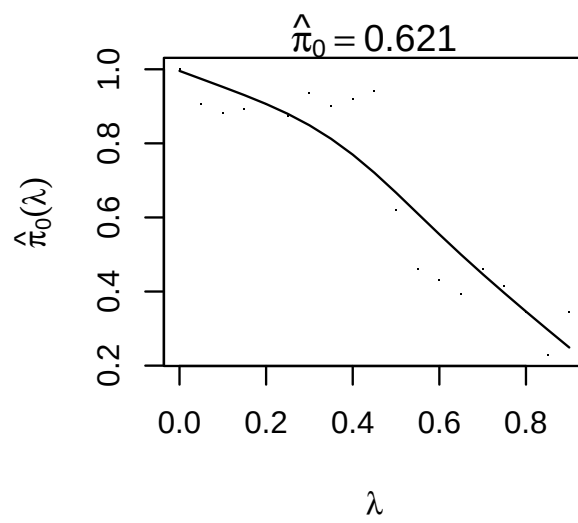


Figure 3: *Q-Q Plot and 80% Guide Rails. Portoscuso example*

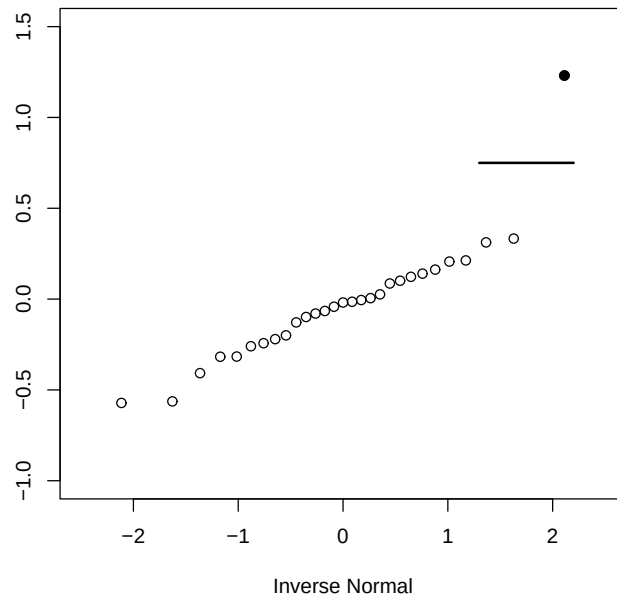


Figure 4: *Q-Q Plot and 80% Guide Rails. La Maddalena example.*

