

Title:

Fitting the distributions of length of stay by parametric models

Running head:

Fitting distributions of length of stay

Authors:

Alfio Marazzi¹, PhD, Fred Paccaud¹, MD, Christiane Ruffieux¹, Mathematician,
Claire Beguin², MD

Affiliations:

¹ Institute of Social and Preventive Medicine, School of Medicine, University of Lausanne, Switzerland

² Center for Medical Informatics, Cliniques Universitaires Saint Luc, Brussels, Belgium

Correspondence and reprint address:

Alfio Marazzi
Institute of Social and Preventive Medicine,
Rue du Bugnon 17
CH-1005 Lausanne
Switzerland

Key words:

Length of Stay (LOS); Diagnosis Related Groups (DRG); Hospital Casemix; Length of Stay Distribution; Trimming Rules; Asymmetric Distributions; Outliers; Parametric Models; Robust Statistics; M-estimates.

Abstract

OBJECTIVES. The purpose of this study is to assess the adequacy of three widely used models - Lognormal, Weibull, and Gamma - for describing the distribution of length of stay (LOS). This is a fundamental step in the development of outliers resistant (robust) methods for the statistical analysis of this kind of data, where the main objective is to determine measures of average and total resource consumption of groups of patients. Current practice uses several types of trimming rules, many of which are based on the Lognormal model, although theoretical and experimental bases is still insufficient.

METHODS. The three models have been adjusted using robust procedures based on M-estimators to about 5 million stays grouped by Diagnosis Related Groups (DRGs): the resulting 3,279 samples were collected in five European countries during 3 years.

RESULTS. Most of the distributions observed could be fitted with one of these models. The descriptions provided by the Gamma and the Weibull models were similar, and the Gamma model could be omitted. The casemix description provided by the Lognormal-Weibull family is, for certain countries, significantly better than the one provided by the single Lognormal model. Often, for a given DRG and a given country, LOS distributions could be described with the same model over several years. A given DRG, however, usually had to be described by means of different models for different countries.

CONCLUSIONS. Practical and conceptual consequences of the results are discussed. They can be extended to the analyses of other consumption variables used in health services.

Statistical procedures for casemix description, including current rules of trimming, should be improved by means of more flexible families of models.

KEY WORDS: length of stay (LOS); diagnosis-related groups (DRG); hospital casemix; length of stay distribution; trimming rules; asymmetric distributions; outliers; parametric models; robust statistics; M-estimates.

Length of stay (LOS) is an easily available indicator of hospital activity and is used for various purposes, such as management of hospital care^{1 2 3 4 5 6 7 8 9 10 11 12}, quality control^{13 14 15 16 17}¹⁸, appropriateness of hospital use^{19 20 21 22 23}, and hospital planning²⁴. Many hospital casemix schemes, such as Diagnosis-Related Groups (DRGs), have been partly built to get homogeneous LOS, because LOS is considered as a reasonable proxy of resource consumption²⁵.

One of the main objectives is to estimate and predict the total resource consumption of groups of patients, e.g., DRGs. Since the total is a multiple of the mean, the natural computation is the arithmetic mean. Unfortunately, LOS distributions are skewed and contain outliers, making the use of this simple statistic questionable. To overcome these problems, various rules of trimming, some of them based on the Lognormal model, are currently used in practice. The purpose of this paper is to investigate the adequacy of this model as a description of a large variety of samples, and to assess the need of extending the modelling tool.

Background

Typical LOS distributions are skewed (asymmetric) and contain outliers. Both of these features are well known and serious hindrances for the use of the arithmetic mean as an estimate of the expected LOS. Skewness is characterised by a long-sided tail (generally toward high values in the case of LOS); it limits the use of inference techniques based on the normality assumption (e.g., confidence intervals and tests for means). Outliers are values markedly different from most others: when a small number of these values are observed, the sample mean can be much larger than when none is observed. Because the values and the frequency of outliers typically fluctuate from sample to sample, the mean and the related inferences are very unreliable, as shown in the following example.

Example 1. Figure 1 shows the LOS histogram of DRG 35 in 1988, for 315 stays in Belgium (BE, Figure 1a) and for 32 stays in Switzerland (CH, Figure 1b). The arithmetic means are 7.9 (BE) and 25.5 (CH) days. Both distributions contain outliers (not shown on the histograms, which are truncated at LOS=50 days); for example, the Swiss data contain two stays of 374 and 198 days. Removing these two outliers reduces the Swiss mean to 8.1 days (instead of 25.5). The usual t-test for comparing means (which is inappropriate for asymmetric distributions, but nevertheless often used) has a one sided attained significance level (ASL = probability that the test statistic is smaller than its observed value) of 0.999 on the complete data set (i.e., the Swiss mean is significantly larger than the Belgian mean) and of 0.07 when the two outliers are removed (i.e., the Swiss and the Belgian means are not significantly different). In other words, the test result is completely determined by just two stays.

Figure 1. Distribution of LOS, DRG 35 (Other disorders of the nervous system w/o cc), 1988

Because a large number of DRGs must be analysed routinely, automatic procedures are needed for conveniently treating skewness and outliers. A common practice is to remove outliers according to some statistical trimming rules *. Various options are used: trim stays beyond a selected percentile of the data distribution, or trim stays whose LOS is longer than three times the mean LOS²⁶, trim stays whose LOS is larger than $\mu + k\sigma$ (where μ and σ are estimates of the mean and the standard deviation on the log scale and k is a given constant, e.g., $k=3$ for Medicare^{27 28}), or trim stays whose LOS is larger than $q_3 + h(q_3 - q_1)$, where q_1 and q_3 are the first and third sample quartiles of log-LOS and h is a given constant²⁹. The arithmetic mean is then computed on the data remaining after trimming.

Many of these trimming rules are based on the assumption that LOS is distributed according to the Lognormal model. Indeed, the purpose of the logarithmic transformation is to "normalize" the LOS distribution and the constants k and h are chosen according to the expected frequency of data that disagree with the normal model and can, therefore, be trimmed. Unfortunately, a variety of distribution patterns are usually observed among DRGs^{30 31}, and the logarithmic transformation is not always appropriate. In many cases, the histogram is still markedly asymmetric even after the logarithmic transformation. In this case, the trimming rules mentioned above are still robust but not well tailored for the observed distribution pattern.

There are two ways to overcome this problem: look for another symmetry creating transformation, or look for another model to describe the distribution; both ways require the development of new statistical procedures. The purpose of this study was to identify other models to describe the LOS distributions. To this end, new robust procedures based on M-estimators^{32 33 34} were used. The basic properties of these estimators have been thoroughly studied and published, and several programs can be found in public domain (see "Methods", below). By means of M-estimators, an adequate, but not necessarily Lognormal, parametric model could be fitted to the majority of the data. The fitting procedure could be completed with robust inference techniques, as shown in the following example.

Example 2. The density functions drawn with solid lines in Figure 1 have been obtained by fitting a Gamma distribution to each set of the data using a procedure based on M-estimates (for the purpose of comparison, the density functions drawn with thin lines have been fitted to the entire data set by the classical maximum likelihood procedure). The two densities conveniently summarize the pattern of the majority of the data. The means of the fitted densities are 5.9 (BE) and 4.8 (CH) days when the complete data set is used, and 5.9 and 4.4 when the two outliers are removed from the Swiss distribution; thus, the robust estimates of the mean are not substantially influenced by large values. Further, the one-sided ASL of a test procedure for comparing means based on the robust Gamma fits³⁵ are 0.07 (complete data) and 0.05 (outliers removed): thus, the means are not declared as different at the usual 5% level for a two sided test. Moreover, if the Lognormal model is used to fit robustly the same data, the estimated means are 6.6 (BE) and 5.1 (CH), the ASL of the test being 0.08. Thus, the robust tests based on different (but adequate) models are not substantially influenced by the model selection.

To sum up, a variety of distribution patterns (not only the Lognormal one) is usually observed among DRGs; therefore, a meaningful approach to describe a set of samples, while facing the outliers problem, is based on flexible tools that work with a variety of distribution models. Modern robust methods have these qualifications. A fundamental step is, therefore, to identify the models that fit with the usually observed LOS distributions. Three widely used models for asymmetric distributions are considered in this article. Beside the Lognormal model, which is already widely used in this framework, two shorter-tailed models - Weibull and Gamma - are taken into account.

Data

This study used a data base containing 4,758,347 hospital stays from five European countries - Belgium (BE), the Swiss Canton Vaud (CH), Ireland (EI), the Italian region Lombardia (LB), and the United Kingdom (UK) - and three calendar years: 1988, 1989, and 1990. The following abbreviations are used : BE88 to denote the Belgian data of year 1988, CH89 for the Swiss data of year 1989, etc. With this notation, the database includes nine country/year data sets: BE88, CH88, CH89, CH90, EI90, LB89, LB90, UK88, UK90. For each stay, the following characteristics are available: LOS, country, year, and DRG number **. Stays are classified into 478 DRGs. Each DRG/country/year data set is called a *sample* here.

This analysis excluded 232,817 stays of less than one day (i.e., such that LOS=0), 713 samples with less than 20 admissions and 93 samples because of computational difficulties. Table 1 shows, for each country/year, the number stays and samples available for analysis.

Table 1. Number of stays and number of samples, by country and year
--

Methods

In a first step, the Lognormal, the Weibull and the Gamma models were adjusted to each sample, using publicly available programs for M-estimates^{36 37 38} and the robust estimates of the expected LOS were computed as the means of the fitted models.

In a second step, the adequacy of the models for each sample was assessed by means of two tools: a robust version of Cox test for nonnested models^{39 40} and the average trimmed absolute residual (ATAR), an original criterion defined below.

Six Cox tests were performed for each sample, by changing the null and the alternative hypothesis according to the scheme shown in Table 2. The Lognormal distribution was accepted as an adequate sample description if both alternatives - Weibull and Gamma - were rejected; otherwise the Lognormal model is rejected. Similarly, the Weibull model was accepted if both alternatives - Lognormal and Gamma - were rejected, etc. Each test was performed at the 2.5% level; thus the probability of wrongly rejecting a true hypothesis was less than 5%. Note that more than one model can be accepted for a sample.

Table 2. Scheme of hypothesis testing with the Cox test
--

According to these rules (based on six tests per sample), each sample could be allocated to one of the following eight classes:

- | | |
|---|------|
| 1. The Lognormal model alone fits the sample : | L |
| 2. The Weibull model alone fits the sample : | W |
| 3. The Gamma model alone fits the sample : | G |
| 4. Both Lognormal and Weibull models fit the sample : | LW |
| 5. Both Lognormal and Gamma models fit the sample : | LG |
| 6. Both Weibull and Gamma models fit the sample : | WG |
| 7. The three models fit the sample : | LWG |
| 8. None of the models fits the sample : | None |

The *average trimmed absolute residual* (ATAR) is used to improve the selection procedure. ATAR is a crude measure of adequacy aimed to identify the best model if the Cox model either accepts more than one model or rejects all models. The ATAR quantifies the difference between expected and observed (ordered) observations for the middle part of the distribution. It is defined, for sample j and model M , as :

$$\text{ATAR}_j(M) = \frac{1}{0.75 \cdot n_j} \sum |d_{ij}|,$$

where $d_{ij} = x_{[ij]} - \hat{x}_{[ij]}$, $x_{[ij]}$ denotes the i th ordered observation of sample j , $\hat{x}_{[ij]}$ is its expected value according to a given model M , n_j is the sample size, the sum extending over all i that are larger than 0.1 and smaller than 0.85. If the ordered observations in the middle part of the distribution (i.e., those between the 10th and the 85th percentiles) are, on average, close to their expected values according to the model, the ATAR is a small value. In other words, ATAR provides a quantitative measure of model adequacy with a focus on the middle part of the LOS distribution.

For a given sample and a given set S of models, the model with the smallest ATAR is designed as *the best model in S according to the ATAR criterion*. The set S may contain a single family of models (e.g., the Lognormal family) or more than one family (e.g., both the Lognormal and the Weibull families). In the first case, the process of selecting a model from S , according to the ATAR criterion, is called a *single model strategy*; in the second case, a *two-model strategy*. For a given country/year, an overall measure of prediction error (ε) of a strategy based on S , is defined as

$$\varepsilon(S) = \sum n_j \cdot 0.75 \cdot \text{ATAR}(M_j),$$

where j extends over all samples of the given country/year, and M_j denotes the best model in S for sample j .

Results

Adequacy of Samples With the Three Models under Study

Table 3 shows the number of samples for which one, more than one, or none of the models was accepted as an adequate description according to the Cox test. At least one model was accepted for two third of the samples. The Lognormal model showed a good acceptance, as the single best description (679), or as equivalent with other models (223 with Gamma, 5 with Weibull, and 668 with both Gamma and Weibull). On the other hand, a substantial number of samples did not fit with the Lognormal model, but with Gamma alone (81), or Weibull alone (92), or both with Weibull or Gamma but not with the Lognormal model (320).

Overall, the Lognormal model was the single best fit for 21% of samples and the non-Lognormal models (Weibull and Gamma) for 15%; Lognormal and non-Lognormal models may be used equivalently for a further 27% of samples. Thus, considering two models other than the Lognormal one did improve the description of the LOS distribution.

<p><i>Table 3. Distribution of samples across three models (Lognormal=L, Weibull=W, Gamma=G) and their combination, according to the Cox test, by country and year</i></p>

Approximately one third of the samples (1,211 out of 3,279) could not be associated with any models according to the Cox test. A systematic visual inspection suggested three main situations responsible for rejection :

- an early peak in the distribution (generally at LOS=1) combined with a strong concentration over a few consecutive days (e.g., 90% of the stays are smaller than 3 days);

- plurimodality, which might be related to the fact that the DRG system fails to attain LOS homogeneity (if the grouping rules cannot be improved, a mixture of models should be used to describe these samples);
- the large size of samples is, by far, the most frequent explanation. A separate analysis showed that the mean size of the samples for which the three models were rejected is 2,909 stays, while the mean size of samples for which at least one model was accepted ranged between 65 and 1,415 stays. A visual inspection (with probability plots) showed that many of these samples were in fact fairly approximated with one of the models. For example, Figure 2 shows the distribution of a sample (DRG 14 in the United Kingdom, 1990) with 20,635 stays, where none of the models was accepted according to the Cox test; nevertheless, the Lognormal distribution was indeed a fair approximation.

Figure 2. LOS distribution and adjusted Lognormal density, DRG 14 (Specific cerebrovascular disorders except TIA), United Kingdom, 1990

From a practical point of view, the most convenient situation is when each sample is assigned to one single model. However, Table 3 shows that a large proportion of samples could be fitted with more than one model, and many samples could not be fitted because of the large number of stays. Therefore, the ATAR criterion (see "Methods" above) has been used to allocate one single model to each sample. Table 4 shows the distribution of stays among the three models. The Lognormal model showed the best rate of allocation of 52%; Weibull had an allocation rate of 20% and Gamma of 28%. Here again, considering two models other than the Lognormal one did improve the adequacy of the description of the LOS distribution. A separate analysis (not reported here) showed that there is a fair concordance between the Cox test and ATAR criterion: the vast majority of the samples that are allocated to class L (Lognormal alone) by the Cox test were also allocated to class L by the ATAR criterion. Similar observations were made for Weibull and Gamma models.

Table 4. Distribution of samples across three models (Lognormal=L, Weibull=W, Gamma=G), according to the ATAR criterion, by country and year

From Three to Two Distribution Models

Table 3 suggests that the performance of the Weibull and the Gamma models were quite similar: 320 samples fit with both models, against 223 samples which fit equally well with Lognormal and Gamma, and five fit with both Lognormal and Weibull models. This opened the possibility of a further reduction of the number of models, choosing between Weibull and Gamma. One way to check the concordance between Weibull and Gamma was to look at the M-estimates of expected LOS when all 3,279 samples are fitted with both models; Table 5 shows that these estimates were very close. In the scatter diagram on Figure 3, all pairs of M-estimates provided by these two models are displayed: the relation was very strong, confirming that the descriptive performance of Weibull and Gamma are very close. These two models, therefore, are redundant.

Facing the choice between Weibull and Gamma, it is reasonable to choose Weibull; indeed, according to Table 3, Weibull plays the best complementary role with respect to Lognormal, because the number of samples fitting equally well with both these models (five samples) was lower than the number of samples fitting with both Lognormal and Gamma (223). Further, algorithms for fitting the Gamma model were computationally more cumbersome than those for fitting the Weibull model.

Table 5. M-estimates of the mean LOS (days), by country and year

Figure 3. Relationships between mean LOS estimates provided by the Weibull and Gamma models

Classification in Two Models: Lognormal and Weibull

The ATAR criterion provides a simple rule to decide which model fits to a given LOS distribution. Table 6 shows the distribution of the samples between the Lognormal and the Weibull models, by country/year. This table (which should be compared to Table 3 and Table 4) confirms the slight preference for the Lognormal model (1,880/1,399). However, the Weibull model was chosen for a very substantial minority of samples (43%).

Another point worth making is the existence of a "country preference" between these two models. Two thirds of samples (between 57% and 69%) were best fitted with the Lognormal model, except in Lombardia which shows a strong preference for the Weibull model (61% and 66% for the two years considered). This "country preference" is quite stable over several years. A further analysis based on samples from the same countries that provided samples for more than one year gives the following results:

- in Switzerland, 220 samples could be found covering three consecutive years (so-called triplets): the Lognormal model best fits 85 triplets and the Weibull model best fit 24 triplets. Therefore, the Lognormal model is the best fit for three years for 50% of the samples.
- In the United Kingdom, there are 424 samples that covered two years (pairs); the Lognormal model best fit 228 pairs and the Weibull model 76 pairs : 72% of the samples were attributed to the same model over the two years.
- Lombardia has 402 samples that covered two years; the Lognormal model best fit 94 pairs and the Weibull model 207 pairs, giving a stable selection for about 75% of the samples.

Table 6. Distribution of samples across two models (Lognormal or Weibull) according to the ATAR criterion, by country and year

This emphasizes the fact that an ad hoc analysis of each set of data is needed to warrant a good descriptive performance of a set of data.

Analysis of a Casemix: Impact of the Model Selection Strategy

The practical impact of using a single-model compared with a two-model strategy was assessed by means of the overall error ε (see "Methods" above). The relative difference

$$r = (\varepsilon(\{L\}) - \varepsilon(\{L, W\})) / \varepsilon(\{L\})$$

between the prediction errors quantifies the improvement gained by one strategy over the other. Here $\{L\}$ represents the Lognormal family, and $\{L, W\}$ the Lognormal-Weibull family. The r values range from 6 to 15%, except for Lombardia, where the impact is substantially higher (30%): this comes as no surprise since in Lombardia, Weibull is the preferred model. Thus, its inclusion as a complement to Lognormal substantially improves the overall performance of the description.

Another direct comparison is presented in Table 7, where casemix adjusted LOS means were obtained according to the strategies based on $\{L\}$ and $\{L, W\}$ and two current trimming rules. In general, the two-model strategy provided relatively low estimates (lower than the arithmetic mean); this was an expected effect of the robust procedure. In Belgium, Lombardia and Ireland, the estimates provided by the single Lognormal strategy, were higher than the crude arithmetic means. In these countries, the single Lognormal strategy overestimates the right distribution tails; its relevance was, therefore, seriously compromised.

The trimming method of Medicare (T1) provided estimates that are larger than or similar (UK88 and EI) to those obtained with the two-model strategy; the estimates provided by trimming rule T2 were smaller or similar. However, these similarities are related to the specific distribution pattern of each country: any change of this pattern would probably result in a reduction in similarities.

<p><i>Table 7. Estimates of mean LOS, adjusted for casemix differences between countries, by country and year</i></p>

Discussion

This paper is a contribution to the development of the statistical analysis of LOS distributions or other consumption variables in health services. In this field, the need of outliers resistant procedures has long been recognised^{41 42 25} and various trimming rules have been proposed. Most of these rules are based on the Lognormal model, this choice being motivated by the long right tails observed in most LOS distributions. Further, it is appealing to have a single standard transformation such as the logarithmic one, to avoid time-consuming analysis of each sample in order to identify the most appropriate transformation.

As noted in the literature⁴³, however, there is a large variety of distributions patterns of LOS, and many patterns are quite far from the Lognormal model. Conversely, progress in statistical and computing provides new tools for robust model fitting, as well as automated treatment of large sets of samples. In this perspective, this study considered the extension of the current modelling tools and the associated procedures for data analysis. Three models (instead of one) were considered: Lognormal, Weibull, and Gamma. This model extension was motivated by the remark that the (right) tail of many observed LOS distributions tends to become shorter and shorter as long as cost reduction policies continue to produce their expected effects. A former study showed that the Exponential model, which is a particular case of the Gamma model, had insufficient flexibility³⁰.

The study has been based on a large database of international provenance: this warrants a good validity of the results. They can be summarised as follows :

- the Lognormal model was actually the model which fit with the LOS distribution of the majority of samples, according either to the Cox test (Table 3) or to the ATAR criterion (Table 4 and Table 6). This confirms the results of an earlier study³⁰. A large minority of the samples did not fit with the Lognormal model and, moreover, the Lognormal predominance was not observed in all countries (Table 3, Table 4 and Table 6); further, the LOS means based on the Lognormal model were substantially higher than the means obtained by the use of other models (see Table 5 and Table 7): this was due to the heavy tail of the Lognormal distribution which frequently overestimates the real length of the observed tail;

- when, besides the Lognormal model, two other models (Weibull and Gamma) were introduced (both characterised by shorter right tails), there was an improvement of the overall description. The model extension provided an adequate fit for a substantial number of samples, according to the Cox test (see Table 3); this improvement had an impact on the mean estimates, with a substantial reduction in the overall prediction error (Table 7). In one country in this study, the estimates of the LOS mean provided by the Lognormal-Weibull combination was substantially lower (2 days) than Lognormal estimates;
- the Gamma model was quite redundant with the Weibull model; the latter can be retained as a complement to the Lognormal model, with the advantage of being computationally simpler than the Gamma model;
- some countries «preferred» Lognormal, others preferred Weibull. This phenomenon might reflect a systematic difference in hospital practice between regions. In fact, not only the mean values of LOS are likely to vary according to place⁴⁴, but also the pattern of distribution of LOS itself;
- although the majority of samples fit with one model according to the Cox test, there were still a substantial number of samples for which other models should be investigated (Table 3); however, the ATAR criterion proposed in this study allows the convenient allocation of almost all samples to one model;
- for many samples, however, no model was adequate according to the testing procedure. The most frequent explanation, apart from the large size of the samples, was a high proportion of stays for which the LOS is equal to one day, which may correspond to hospitalization transfer between wards. For these situations other models (e.g., discrete models) will have to be studied. In the meantime, the Weibull or the Lognormal model should be selected according to a crude criterion of goodness of fit.

The practical consequence of using a variety of models in place of a single one is the need to develop and implement new techniques for fitting and selecting these models to estimate the mean LOS on their ground, and to make inferences. New rules of trimming asymmetric data modelled as Weibull and Gamma are under development. In this paper, M-estimators were used because their theory is well known thanks to an extensive statistical literature. They can be adapted to different parametric models and completed with inference techniques with a relatively limited effort. For example, the Cox test was adapted to the selection of models and a few tests for comparing means were used for the examples. Although these procedures might be improved, they already can be used following the simple guidelines below :

1. Fit all samples with both the Lognormal and the Weibull models using the available procedures based on M-estimators;
2. Use the ATAR criterion to assign a single model to each sample; the M-estimate of the LOS mean is then automatically obtained from the fitting procedure; these LOS means can be used as a descriptive measure, but not for inference;
3. Assess the quality of the fit with the Cox test; however, a visual inspection (and hence, some subjective judgement) is required for those samples that reject both models; for those samples for which a model is adequate, inferences can be made on the ground of well known properties of M-estimators, including statistical comparison of the means.

Improving statistical methods is a real challenge for the future of LOS analysis or other similar resource variables. A first challenge is the improvement of the statistical monitoring of

health services. Those working on LOS and other variables like costs will have to live for a long time with skewness and outliers^{45 4}. Certain DRGs may be too vague⁴⁶ and many ongoing efforts are aimed at the improvement of the homogeneity of DRGs^{47 48 49}: however, it is unlikely that they will completely eliminate outliers⁵⁰. It should be recognized that the substantial presence of outliers makes data interpretation more difficult, for example, in the analysis of the variation of LOS among hospitals⁵¹. The statistical treatment of large database requires procedures which are both rapid and sound, i.e., based on an explicit theory and a careful assessment. The current method of analyzing data is often poor, with indicators which are meaningless because they lack of statistical grounds⁵². Further, the statistical significance of LOS difference (e.g., for the assessment of the impact of various programs) can be meaningless if the modelling of the LOS distribution is neglected^{53 52}. The problem is still more important when LOS is used as a dependant variable in regression analysis. In fact, the wide variation of LOS transformations and/or trimming procedures which is observed in the literature is an indication that a strong theoretical effort is needed to develop a common approach.

Another point is that outliers can give important information for various aspect of health services management. In that sense, outliers are not only values to be eliminated, but rather values to be taken into consideration for the assessment of various interventions^{54 55 56 52} or the allocation of resources^{11 45 57 58 59}. There are many other developments in the use of outliers, for example, identifying outlier institutions in terms of management⁶⁰ or quality of care⁶¹. In any case, however, there is a need to develop the underlying statistical concepts aimed at identifying these outliers.

A better use of statistical techniques is important to improve the comprehension of the functioning of health services. Robust statistics provides powerful tools, but yet underused, in the field of health care. An important paradigm of robust statistics is that a parametric model can be used to describe the "majority", not the "totality" of cases. More precisely, one assumes that the population is a mixture of a parametric distribution (the model) and an unspecified distribution, called contamination, that describes the outliers. This paradigm opens a more flexible approach to the problem of managing exceptional cases, for example for prospective payment, than currently used rules, where each single stay exceeding some limit deserves a special treatment for outliers. Indeed, a robust model describes the "regular" stay distribution that includes rare but expected (and hence, legitimate) long stays, whereas the contamination describes the "irregular" (exceptional and unexpected) stays due to some kinds of accidents (ranging from errors in the codification process to catastrophic medical situations). Robust methods could be used to characterize the "regular" distribution according to the model as well as the amount of contamination. Prospective allocation could then take advantage of this description: for example, a regular cost for next year (that includes a certain amount of legitimate large stays) could be estimated as the mean of the model fitted to this year data majority. Financial reserve for outliers could be made on the grounds of the amount of contamination. Control and reimbursement then could be based on the amount of stays that disagree with the model at the end of next year.

It may, finally, be noticed that the process of model choice does not need to be based solely on criteria of agreement with data. It also may take into account the agreement with the management target. For example, a short-tailed model (e.g., some member of the Weibull class) is best suited to describe and stimulate a cost reduction target than a long-tailed model (Lognormal).

Acknowledgements

The authors are grateful to their colleagues who kindly provided the data used in this study : Francesca Repetto (Milan, Italy), Miriam M. Wiley (Dublin, Eire), Marie-Christine Closon (Brussels, Belgium), and Phil Anthony (Winchester, UK); relevant remarks were provided by Prof. Roger France (Brussels). Alex Randriamiharisoa (Lausanne) helped in programming. This study has been financed by the Swiss National Foundation for Scientific Research, Grant No. 32-39614.93

End Notes

* The other strategy is to use fixed trimming, i.e., a cut-off point beyond which the LOS is considered as impossible for some clinical reason or other common sense aspects.

** Distributions and information about specific DRGs are available from the first author.

References

- ¹ Fernow LC, McColl I, Mackie C. Firm, patient, and process variables associated with length of stay in four diseases. *BMJ* 1978;1:556
- ² Carstairs V, Watkins G, Young P. The use of surgical beds: variations between consultants in duration of stay for selected operations. *Health Bull* 1978;36:162
- ³ Schwartz RJ, Jacobs LM, Yaezel D. Impact of pre-trauma center care on length of stay and hospital charges. *J Trauma* 1989;29:1611
- ⁴ Conrad D, Wickizer T, Maynard C, et al. Managing care, incentives, and information : an explanatory look inside the "Black Box" of hospital efficiency. *Health Services Res* 1996;31:235-59
- ⁵ Reder VA, Fineberg HV, et al. Shorter length of stay for simple cholecystectomy: cost-effectiveness of alternative strategies. *Med Care* 1983;21:745
- ⁶ Clendenning MK, Wolfe H, Shuman LJ, et al. The effect of a target date based utilisation review program on length of stay. *Med Care* 1976;9:751
- ⁷ Roach JA, Tremblay LM, Bowers DL. A preoperative assessment and education program : implementation and outcomes. *Patient Educ Counselling* 1995;25:83.
- ⁸ Robinson JC, Luft HS, McPhee SJ, Hunt SS. Hospital competition and surgical length of stay. *JAMA* 1988;259:696
- ⁹ Studnicki J. Differences in length of stay for Medicaid and Blue Cross patients and the effects of intensity of services. *Publ Health Rep* 1979;94:438.
- ¹⁰ Evans JH, Hwang YM, Nagarajan N. Physicians' response to length of stay profiling. *Med care* 1995;33:1106
- ¹¹ Bernard AM, Hayward RA, Rosevear J, et al. Comparing the hospitalizations of transfer and non-transfer patients in an academic medical center. *Acad Med* 1996;71:262-6
- ¹² Flamer HE, Christophidis N, Margetts C, et al. Extended hospital stays with increasing age: the impact of an acute geriatric unit. *Med J Australia* 1996;164:10-3
- ¹³ Epstein AM, Bogen J, Dreyer P, Thorpe KE. Trends in length of stay and rates of readmission in Massachusetts: implications for monitoring quality of care. *Inquiry* 1991;21:19
- ¹⁴ Riley G, Lubitz J, Gornick M, et al. Medicare beneficiaries : adverse outcomes after hospitalisation for eight procedures. *Med Care* 1993;31:921
- ¹⁵ Knaus WA, Wagner DP, Zimmermann JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med* 1993;118:753-61
- ¹⁶ Cleary PD, Greenfield S, Mulley AG, et al. Variations in length of stay and outcomes for six medical and surgical conditions in Massachusetts and California. *JAMA* 1991;266:73-9
- ¹⁷ Painter LM, Dudjak LA, Breiner K, et al. Abdominal aortic aneurysm pathway: outcome analysis. *J Vascular Nursing* 1995;13:101-5
- ¹⁸ Korda H. Utilization review for Medicaid DRGs systems : practice, innovation, and lessons of experience. *Am J Med Qual* 1994;9:54-67
- ¹⁹ Fortney JC, Booth BM, Smith GR. Variation among VA hospitals in length of stay for treatment of depression. *Psychiatric Services* 1996;47:608-13
- ²⁰ Mushlin AI, Black ER, Connolly CA, et al. The necessary length of stay for chronic pulmonary disease. *JAMA* 1991;266:80-3
- ²¹ Chen E, Naylor D. Variation in hospital length of stay for acute myocardial infarction in Ontario, Canada. *Med Care* 1994;32:420
- ²² Sharp JW, Coleman E, Starling N, et al. Hospital utilization for AIDS: are all hospital days necessary? *Quality Rev Bull* 1991;17:113

- ²³ Selker HP, Beshansky JR, Pauker SG, Kassirer JP. The epidemiology of delays in a teaching hospital : the development and use of a tool that detects unnecessary hospital days. *Med Care* 1989;27:112
- ²⁴ Lagoe RJ. A community-based analysis of regional differences in hospital stay by DRGs. *Inquiry* 1986;23:183
- ²⁵ Lave JR, Leinhardt S. An evaluation of a hospital stay regulatory mechanism. *Am J Public Health* 1976;66:959
- ²⁶ Casemix funding for public hospitals 1994-95. Melbourne: Victorian Department of Health and Community Services, 1994
- ²⁷ Schweiker RS. Report to the Congress: Hospital prospective payment of Medicare. Washington, DC: US Govt printing office, DHHS, 1983
- ²⁸ Department of Health and Human Services. Medicare program. Prospective payment of Medicare inpatients: Interim final rule. *Federal Register* 1983; 48:39752-890.
- ²⁹ Beguin C, Closon MC, Roger FH. Advances in DRGs data pooling in Europe: results from the Hoscom Project in relation to outliers. Paper presented at the second EURODRG Workshop, Dublin, 24-25 April 1991.
- ³⁰ Ruffieux C, Marazzi A, Paccaud F. Exploring models for the LOS distribution. *Soz Präventivmed* 1993;38:77-82
- ³¹ Marazzi A, Ruffieux C, Paccaud F. Trimming rules and M-estimates for the expected length of stay in Diagnosis Related Groups. Proceedings of the 9th International PCS/E Working Conference, Muenchen 15-18 September 1993.
- ³² Huber PJ. Robust statistics. John Wiley & Sons, Inc, 1980
- ³³ Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust statistics. The approach based on influence functions. John Wiley&Sons, Inc, 1986
- ³⁴ Marazzi A, Ruffieux C. Implementing M-estimators of the Gamma distribution. In Rieder E (Ed), *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Springer-Verlag, 1996.
- ³⁵ Marazzi A. Estimating and testing the means of asymmetric distributions using M-estimators. Submitted.
- ³⁶ Marazzi A, Randriamiharisoa A. S-plus functions for M-estimates of the Gaussian and the Lognormal distributions. Technical Report. Institut Universitaire de Médecine Sociale et Préventive, Bugnon 17, CH-1005 Lausanne. (The report and the programs are available at the Internet address www.hospvd.ch/iump).
- ³⁷ Marazzi A, Randriamiharisoa A. S-plus functions for M-estimates of the Gamma distribution. Technical Report. Institut Universitaire de Médecine Sociale et Préventive, Bugnon 17, CH-1005 Lausanne. (The report and the programs are available at the Internet address www.hospvd.ch/iump).
- ³⁸ Marazzi A, Randriamiharisoa A. S-plus functions for M-estimates of the Weibull distribution. Technical Report. Institut Universitaire de Médecine Sociale et Préventive, Bugnon 17, CH-1005 Lausanne. (The report and the programs are available at the Internet address www.hospvd.ch/iump).
- ³⁹ Cox DR. Further results on test of separate families of hypotheses. *J R Statistic Soc B* 24, 1962, 406-24
- ⁴⁰ Victoria-Feser MP. A robust test for non-nested hypotheses. *J R Statistic Soc B* 59, 1997, 715-727
- ⁴¹ Gustafson R. Length of stay: prediction and explanation. *Health Serv Res* 1968;3:12-34
- ⁴² Lave JR, Lave LB. The extent of role differentiation among hospitals. *Health Serv Res* 1971;6:15-38
- ⁴³ Dussaucy A, Viel JF. Typologie des Groupes Homogènes de Malades de la base de référence française en fonction de leur distribution de durée de séjour.
- ⁴⁴ List ND, Fronczack NE, Gottlieb SH, Baker RE. A cross-national study of differences in length of stay of patients with cardiac diagnoses. *Med Care* 1983;21:519-30
- ⁴⁵ Siegel JH, Shafi S, Goodarzi S, Dischinger PC. A quantitative method for cost reimbursement and length of stay quality assurance in multiple trauma patients. *J Trauma* 1994;37:928-37
- ⁴⁶ Berki SE, Ashcraft MLF, Newbrander SC. Length of stay variations within ICDA-8 DRGs. *Med Care* 1984;22:126-42
- ⁴⁷ Freeman JL, Fetter RB, Park H, et al. Diagnosis-Related Group refinement with diagnosis- and procedure-specific comorbidities and complications. *Med Care* 1995;33:765

- ⁴⁸ Hicks N, Kammerling R. The relationship between a severity of illness indicator and mortality and length-of-stay. *Health Trends* 1993;25:65-8
- ⁴⁹ Quantin C, Mathy C, Moreau T, et al. Structural and conjunctural compensation method for hospital budgetary allocation on the basis of DRGs. *Medinfo* 1995;8 Pt 1:537-40
- ⁵⁰ Criner GJ, Kreimer DT, Tomaselli M, et al. Financial implications of non-invasive positive pressure ventilation. *Chest* 1995;108:475-81
- ⁵¹ Fortney JC, Booth BM, Smith GR. Variation among VA hospitals in length of stay for treatment of depression. *Psychiatric Services* 1996;47:608-13
- ⁵² Bernard AM, Hayward R, Anderson JE, et al. The Integrated Inpatient Management Model : lessons from managed care. *Med Care* 1995;33:663
- ⁵³ Cleary PD, Greenfield S, Mulley AG, et al. Variations in length of stay and outcomes for six medical and surgical conditions in Massachusetts and California. *JAMA* 1991;266:73-9
- ⁵⁴ Philipson EH, Curry SL. Quality assurance: measuring its effect on a busy obstetric service. *Obstetrics & Gynaecology* 1994;83:131-3
- ⁵⁵ Clendenning MK, Wolfe H, Shuman LJ, et al. The effect of a target date based utilisation review program on length of stay. *Med Care* 1976;9:751
- ⁵⁶ Rubin JW. Video-assisted thoracic surgery: the approach of choice for selected diagnosis and therapy. *Eur J Cardiothoracic Surgery* 1994;8:431-5
- ⁵⁷ Mossman D, Songer DA, Baker DG. Predicting length of children's psychiatric hospitalisations : an "ecological" approach. *Quality Rev Bull* 1991;17:269
- ⁵⁸ Breckwoldt WL, Mackey WC, O'Donnell TF. The economic implications of high-risk abdominal aortic aneurysms. *J Vasc Surgery* 1991;13:798
- ⁵⁹ Criner GJ, Kreimer DT, Tomaselli M, et al. Financial implications of non-invasive positive pressure ventilation. *Chest* 1995;108:475-81
- ⁶⁰ Fortney JC, Booth BM, Smith GR. Variation among VA hospitals in length of stay for treatment of depression. *Psychiatric Services* 1996;47:608-13
- ⁶¹ Park RE, Brook RH, Kosecoff J, Keesey J, Rubenstein LV, Keeler EB, Kahn KL, Rogers WH, Chassin MR. Explaining variations in hospital death rates: randomness, severity of illness, quality of care. *JAMA* 1990;264:484-90

Table 1. Number of stays and number of samples, by country and year

<i>Country and year</i>	<i>Number of stays available</i>	<i>Number of stays included</i>	<i>Number of samples included</i>
<i>BE 88</i>	90872	87092	406
<i>CH 88</i>	29112	27432	249
<i>CH 89</i>	30127	28901	256
<i>CH 90</i>	30457	28494	250
<i>EI 90</i>	214896	193785	398
<i>LB 89</i>	305221	304772	404
<i>LB 90</i>	1092867	1080719	439
<i>UK 88</i>	670008	541955	427
<i>UK 90</i>	2294787	2087650	450
<i>all</i>	<i>4758347</i>	<i>4380800</i>	<i>3279</i>

BE is Belgium, CH is Switzerland (Canton Vaud), EI is Ireland, LB is Lombardia, UK is United Kingdom, 88 is year 1988, etc. For each country/year, the number of stays available in the data base and the number of stays and samples included in the analysis are reported

Table 2. Scheme of hypothesis testing with the Cox test

<i>Model under test for adequation with data</i>	<i>Null hypothesis</i>	<i>Alternative hypothesis</i>
<i>Lognormal (L)</i>	L	G
	L	W
<i>Weibull (W)</i>	W	G
	W	L
<i>Gamma (G)</i>	G	L
	G	W

Table 3. Distribution of samples across three models (Lognormal=L, Weibull=W, Gamma=G) and their combination, according to the Cox test, by country and year

	<i>BE88</i>	<i>CH88</i>	<i>CH89</i>	<i>CH90</i>	<i>EI90</i>	<i>LB89</i>	<i>LB90</i>	<i>UK88</i>	<i>UK90</i>	<i>all</i>
<i>L</i>	113	46	47	34	94	68	75	109	93	679
<i>W</i>	6	2	5	6	7	22	29	5	10	92
<i>G</i>	9	2	1	2	4	18	18	13	14	81
<i>LW</i>	0	0	0	1	2	0	0	0	2	5
<i>LG</i>	42	31	29	32	28	24	17	15	5	223
<i>WG</i>	38	12	18	23	41	67	61	34	26	320
<i>LWG</i>	104	106	101	99	91	73	34	42	18	668
<i>None</i>	94	50	55	53	131	132	205	209	282	1211
<i>all</i>	406	249	256	250	398	404	439	427	450	3279

Table 4. Distribution of samples across three models (Lognormal=L, Weibull=W, Gamma=G), according to the ATAR criterion, by country and year

	<i>BE88</i>	<i>CH88</i>	<i>CH89</i>	<i>CH90</i>	<i>EI90</i>	<i>LB89</i>	<i>LB90</i>	<i>UK88</i>	<i>UK90</i>	<i>all</i>
<i>L</i>	255	149	139	138	214	145	124	278	277	1719
<i>W</i>	68	43	43	42	73	92	126	77	78	642
<i>G</i>	83	57	74	70	111	167	189	72	95	918
<i>all</i>	406	249	256	250	398	404	439	427	450	3279

Table 5. M-estimates of the mean LOS (days), by country and year

	<i>BE88</i>	<i>CH88</i>	<i>CH89</i>	<i>CH90</i>	<i>EI90</i>	<i>LB89</i>	<i>LB90</i>	<i>UK88</i>	<i>UK90</i>
<i>Lognormal (L)</i>	10.9	10.3	9.9	9.7	6.6	11.3	11.4	6.4	9.2
<i>Weibull (W)</i>	9.2	9.3	8.9	8.7	5.6	9.6	9.2	5.4	7.2
<i>Gamma (G)</i>	9.4	9.5	9.1	8.9	5.7	9.7	9.3	5.4	7.2

Table 6. Distribution of samples across two models (Lognormal or Weibull) according to the ATAR criterion, by country and year

	<i>BE88</i>	<i>CH88</i>	<i>CH89</i>	<i>CH90</i>	<i>EI90</i>	<i>LB89</i>	<i>LB90</i>	<i>UK88</i>	<i>UK90</i>	<i>all</i>
<i>Weibull</i>	137	81	97	93	168	247	290	130	156	1399
<i>Lognormal</i>	269	168	159	157	230	157	149	297	294	1880
<i>total</i>	406	249	256	250	398	404	439	427	450	3279

Table 7. Estimates of mean LOS, adjusted for casemix differences between countries, by country and year

<i>Ways to compute the arithmetic mean</i>	<i>BE88</i>	<i>CH88</i>	<i>CH89</i>	<i>CH90</i>	<i>EI90</i>	<i>LB89</i>	<i>LB90</i>	<i>UK88</i>	<i>UK90</i>
<i>crude, all data</i>	10.06	7.89	7.54	6.88	6.40	9.95	9.52	6.48	9.83
<i>...robust, using the Lognormal model</i>	10.49	7.50	7.41	6.79	6.70	10.51	10.55	6.46	9.32
<i>... after trimming according to T1 (see below)</i>	9.89	7.49	7.35	6.82	6.20	9.69	9.30	6.20	9.05
<i>... after trimming according to T2 (see below)</i>	9.48	7.02	6.93	6.41	5.88	9.40	9.09	5.75	8.00
<i>... robust, using the Lognormal and Weibull models</i>	9.47	7.07	6.99	6.46	6.15	9.29	8.94	6.15	8.59

T1 : $\log(t_i) = \mu \pm 3\sigma$ where μ is the geometric mean and σ the square root of the variance of the log (LOS);
T2 : $\log(t_1) = \log(q_1) - 1.15(\log(q_3) - \log(q_1))$ and $\log(t_2) = \log(q_3) + 1.15(\log(q_3) - \log(q_1))$, where q_1 and q_3 are the first and the third sample quartiles.

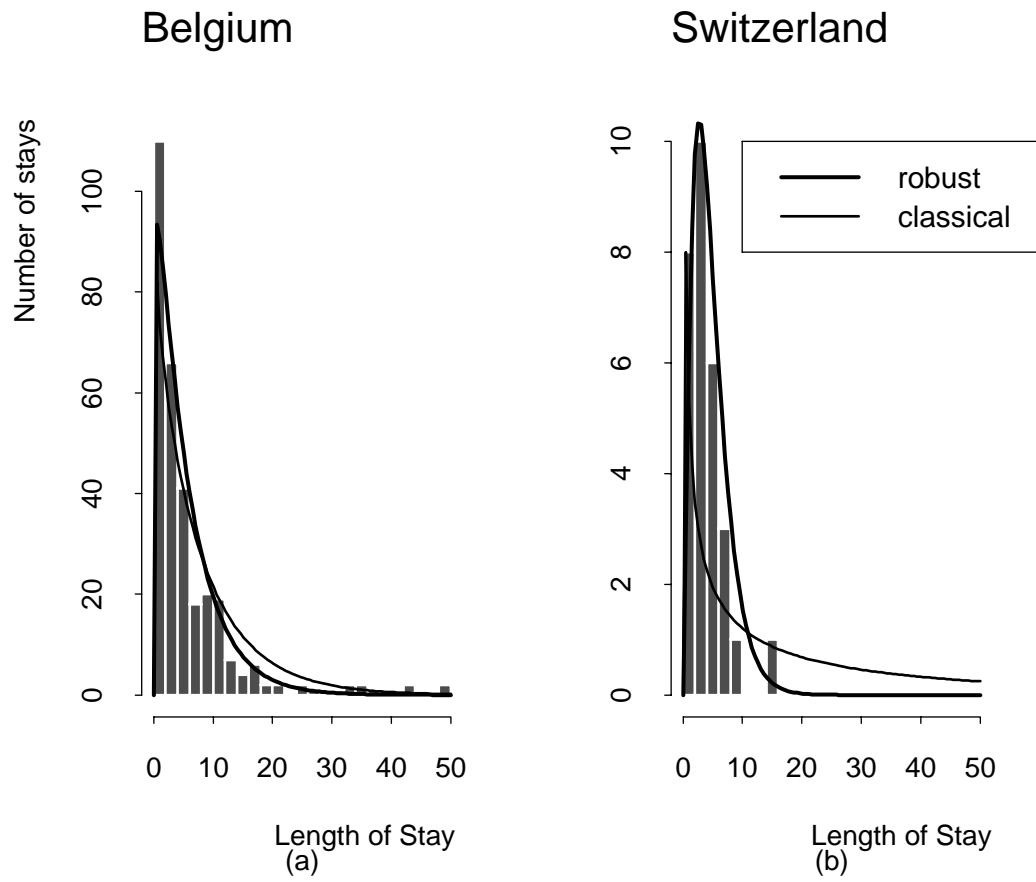


Figure 1 : Distribution of LOS, DRG 35 (Other disorders of the nervous system w/o cc), 1988

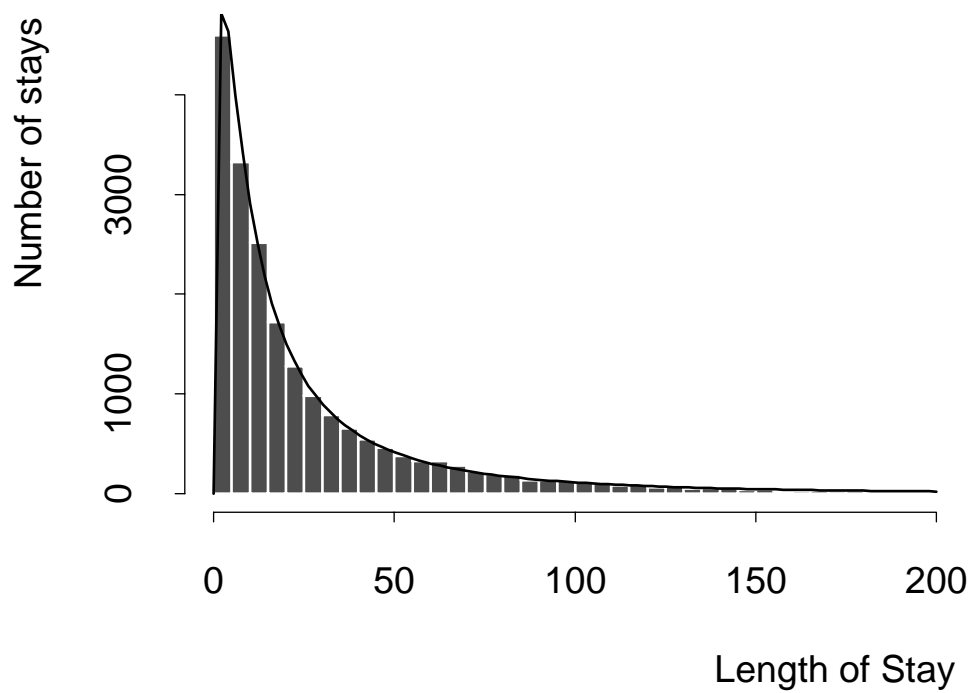


Figure 2 : LOS distribution and adjusted Lognormal density, DRG 14 (Specific cerebrovascular disorders except TIA), United Kingdom,1990

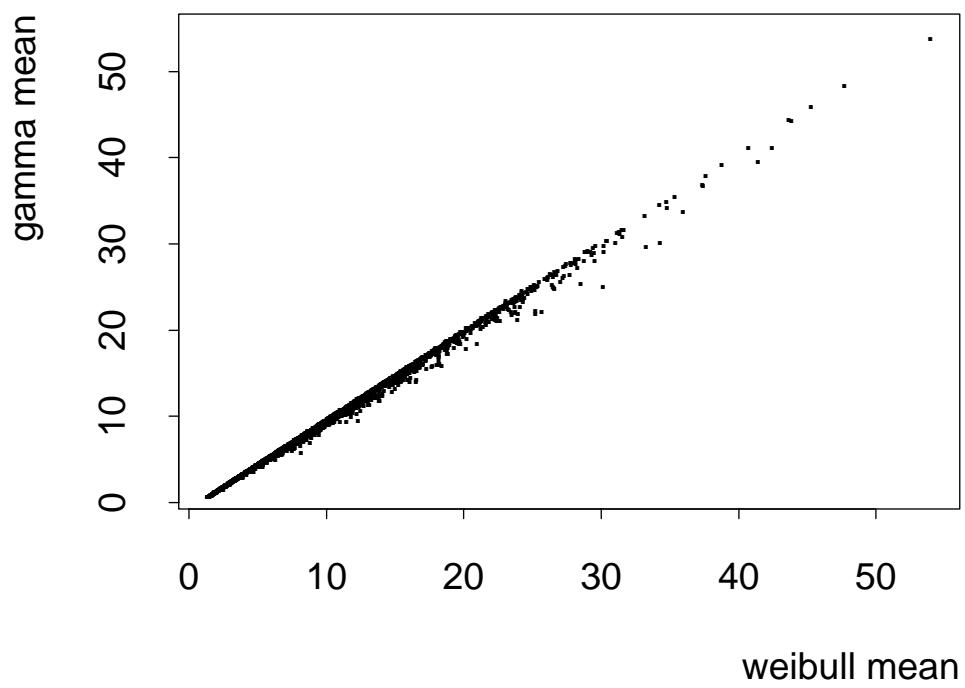


Figure 3 : Relationships between mean LOS estimates provided by the Weibull and Gamma models