Chapter 5

Statistical Analysis

In this chapter the statistical analysis of different measurement sets is carried out. After the first inquiries about the short and long-term characteristics of the process and about the standard deviation of the measurements, the probability density function is compared to the normal distribution. At the end, an empirical fitting curve is performed to enable the numerical calculation of probabilistic estimates.

5.1 Hypothesis on the process

The statistical analysis of the measurement data sets has been realized under the following hypothesis:

1. During short time periods in which the number of pieces in the scale remains still and in absence of induced ground vibrations, measurement reading is considered as a *stationary and homoscedastic* statistical process. This means that its variance and mean value are considered to be constant and independent.

2. It is accepted that the variance can undergo slow variations due to a correlation with the load on the scale or to the influence of the own vibrations of the structure. This hypothesis will be reviewed later at the end of this thesis and is based on the study of mechanical vibrations by Mohd Syazrulazwan Mohd Zin [2].

3. After each event, the process experiments a short transient period [2] after which it returns to be stationary around a new mean value and keeping its variance.

In conclusion, the properties of the process will depend on the term of time in which we are observing it. The difference between the short term and long term characteristics (first and second hypothesis respectively) will lead us to a choice about the kind of test to be used. As it will be shown in chapter 6, applying the hypothesis of stationarity without taking into account the second hypothesis will result in an inefficient test. This will be solved later in the same chapter by introducing the variability of the variance into the main structure of the algorithm.

5.2 Data selection and analysis of the standard deviation

The statistical analysis of the measurements starts with the study of the standard deviation of the process. Therefore 14706 measurements (9 minutes and 4 seconds) were taken in the laboratory in different days and with different load levels of the scale. Then, 29 sets of 51 samples corresponding to stationary situations were selected.

First, the standard deviation is calculated for each set of measurements and then their mean value μ_σ (using Fisher's theorem) and their standard deviation σ_σ. Results are shown in table 5.1. With these data, a 95% confidence interval for σ can be calculated as follows:

$$\boxed{\begin{aligned} I_{0.95}(\sigma) &= \mu_\sigma \pm 1.96\sigma_\sigma = 194.3683 \pm 72.3522 \\ &= [122.0161 \quad \dots \quad 266.7205]\,. \end{aligned}} \tag{5.1}$$

This interval extends ±37% over its central value, so the total relative size of the range of possible values is near to a 74% of the mean value. This means a quite large range of possible values (low accuracy) for σ, and will hinder the choosing of adequate parameters for the test.

This absence of homogeneity of the measurements is on account of the already mentioned long term changes on the value of σ, thus, it is supported by the second hypothesis. A prove of this can be obtained by calculating the mean value and standard deviation of s_i in a short term approach and comparing them with results in (5.1). For example, considering just the first five data sets, that were taken the same day and with similar load levels, we obtain:

$$\begin{aligned} \hat{\mu}_\sigma &= \frac{1}{N \cdot c_2(N)} \sum_{i=1}^{5} s_i = 176.9408 \\ \hat{\sigma}_\sigma &= std(s_i) = 23.9978 \end{aligned} \tag{5.2}$$

A 95% confidence interval for this case is: $I_{0.95} = 176.9408 \pm 47.0357$. This means a range that is smaller than a 54% of the mean value, which is much more tight than the 74% in the long term study.

5.3 Comparison with the normal distribution

Looking for an easy way to make guesses and probability calculations, it is usual to consider a *normal distribution* for the statistical process. Though, this approach is not suitable for every real distribution and its applicability must be tested previously. In this section the results and the procedure on applying the *"χ^2 Goodness of fit test"* are shown.

Table 5.1: Standard Deviation of selected measurement sets

set	$std(\boldsymbol{x})$	set	$std(\boldsymbol{x})$	set	$std(\boldsymbol{x})$	set	$std(\boldsymbol{x})$
1	159.6377	**9**	139.2466	**17**	165.1315	**25**	210.9305
2	205.7948	**10**	155.1914	**18**	161.7695	**26**	250.6496
3	153.6584	**11**	187.6290	**19**	232.3890	**27**	197.0178
4	158.0584	**12**	188.3050	**20**	214.7668	**28**	193.5254
5	194.4694	**13**	248.4934	**21**	182.6947	**29**	206.3681
6	154.8957	**14**	162.2526	**22**	181.2536	**30**	-
7	140.0816	**15**	300.9723	**23**	170.3749	**31**	-
8	212.1855	**16**	199.6762	**24**	225.8919	**32**	-

Average value $\boldsymbol{\mu_\sigma} = \mathbf{194.3683\,g}$; Standard deviation $\boldsymbol{\sigma_\sigma} = \mathbf{36.9144\,g}$

The *"χ^2 Goodness of fit test"* (or just *"χ^2 test"*) is one of the most extended tests to check out the hypothesis of a certain distribution. Its fundamental is the fact that the difference between the ideal *probability density function* (PDF) and the real relative frequency of the data in any specific range value in the total range of the measurements corresponds to a normal null-centered distribution (eq. (5.3)). This allows us to establish bounds for the summation of the squared differences within a determined significance level and to use these bounds to determine if the data correspond to the tested ideal PDF.

$$\chi_n^2 = \sum_{i=1}^{N} \left(f_{id}(\hat{x}_i) - \hat{f}_i \right)^2$$

$$f_{id}(\hat{x}_i) - \hat{f}_i = N(0,1) \tag{5.3}$$

where $\hat{x}_i$ is the normalized value of the center of each segment i and $\hat{f}_i$ is the real relative frequency of data into segment i.

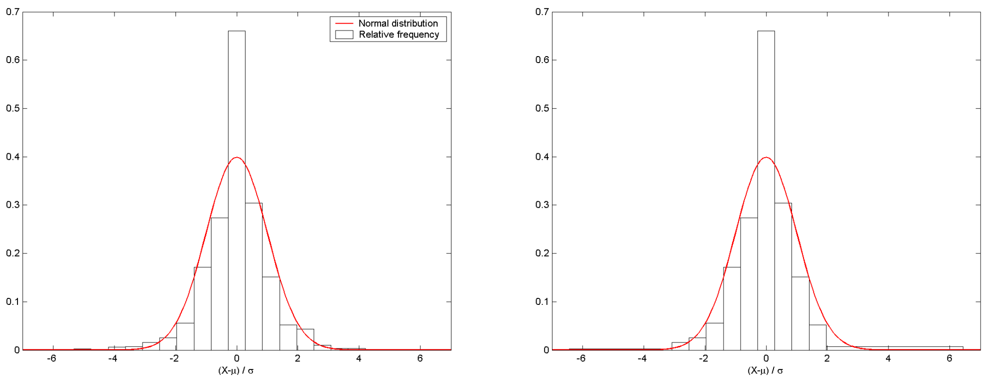
One of the disadvantages of this test is its dependency on the chosen segments. Anyway, it is almost guaranteed to be reliable under the following conditions:

1. The number of bins must be higher than 5.
2. Data must be segmented in such way that there are not empty bins.

In this thesis two different segmentations of 11 segments have been used and both bring similar results.

At first, 1497 of the data used in the analysis of the standard deviation are selected and distributed among 20 equal-spaced segments along the range of values they take obtaining the *histogram* on figure 5.1.*a*. Since empty bins are obtained with this partition, some of them at extremes are joined resulting in an 11 non-equal-spaced bins histogram shown on figure 5.1.*b*.

Figure 5.1: *Two data histograms compared to the normal distribution*



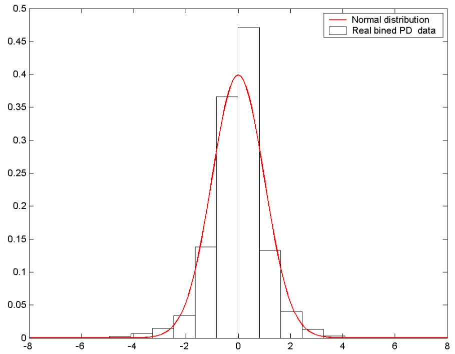
[a]Figure 5.1.*a*

[b]Figure 5.1.*b*

[a]Null-centered segmentation using 23 equal-spaced bins

[b]Same bins are used, but the first 6 and last 8 bins are joined in this case

Figure 5.2: 11 equal-spaced bins data histogram



This has been repeated using directly 11 equal-spaced bins, as shown in figure 5.2. In both cases, the χ^2 test (see table 5.2) is much higher than the lowest limit. In conclusion, the χ^2 test gives negative results.

In view of these results, it is decided to look for any empirical approximation function to the probability density function, that allows the probability estimation and the calculation of the test parameters.

5.4 Empirical adjustment of the probability density function

The adjustment starts as well with the realization of the binned distribution of data relative frequency. This time nineteen equal-spaced segments were chosen of whom the one at the middle is null-centered. The resulting histogram comes out to be quite symmetric, so it is concluded that small asymmetries are just the result of the statistics while the real distribution is considered to be symmetric. Thus, data are changed taking now for each bin the mean value between the original and

Table 5.2: results of the χ^2 test

Degrees of freedom	χ^2	$\chi^2_{(\alpha)}$ $\alpha = 0.01$	$\chi^2_{(\alpha)}$ $\alpha = 0.05$	H_0
10	350.42	2.558	3.940	Rejected
10 (equal spaced segments)	45787.00	2.558	3.940	Rejected

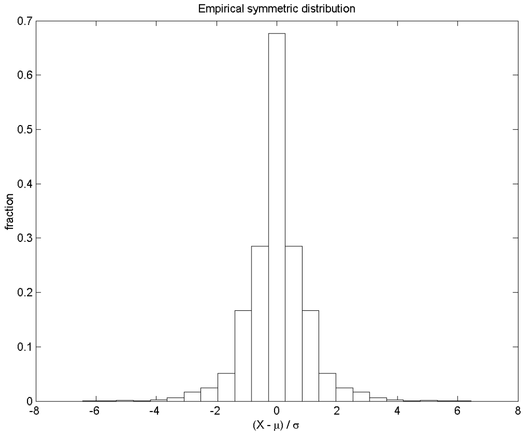


Figure 5.3: Symmetric relative frequency distribution

the one of the respective symmetric bin, resulting in the completely symmetric histogram on figure 5.3.

Like all probability density functions, the fitting curve we are looking for must have an area equal to one under itself. Further more, the area must be equal to the cumulative relative frequency at the end of the segment (right extreme). Thus, the aim is to find at first a curve that steps over the points defined by the cumulative summation of the histogram data. Then, by derivation, the probability density function can be obtained.

Since derivatives increase high frequency errors, it is necessary to be careful while selecting the fitting method for the cumulative summation. Concretely, broken lines must be avoided since their derivatives would have discontinuities. Two methods are proposed here that ensure smooth derivatives:

Low-Pass filter interpolation. The low-pass effect avoids high frequency components and, therefore, broken lines.

Spline interpolation. This method enforces the slope of the curve to be the same at both sides of the input points and, since it is made up of polynomials, its derivative is always smooth.

Both of them produce a curve that fits the empirical data, and both allow smooth derivatives. Graphical results are shown on figures 5.4 and 5.5. Though

Figure 5.4: Low-pass filter fitting curve

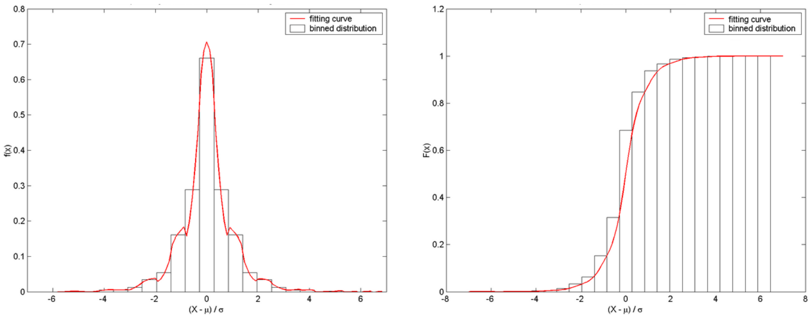
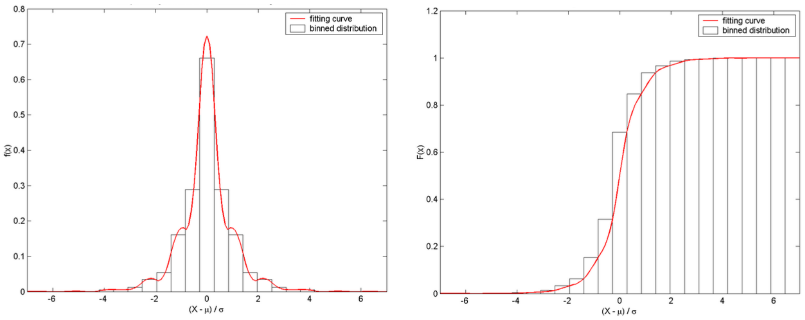


Figure 5.5: Spline fitting curve



they are very similar, the use of *splines* presents some advantages, like the possibility of enforcing a null slope in the *infinite*[1]. Besides, interpolation and derivation with the use of *splines* is much easier and faster once the coefficients are calculated. Therefore, the spline method has been selected.

Figure 5.6 is presented here to justify that both results are very similar, so the argued reasons to choose the *spline* method have importance enough. It shows the difference between both probability distribution curves. This is not higher than 0.25% and, in the range of values $X > 2$, where it is going to be calculated, it is even lower than 0.03%.

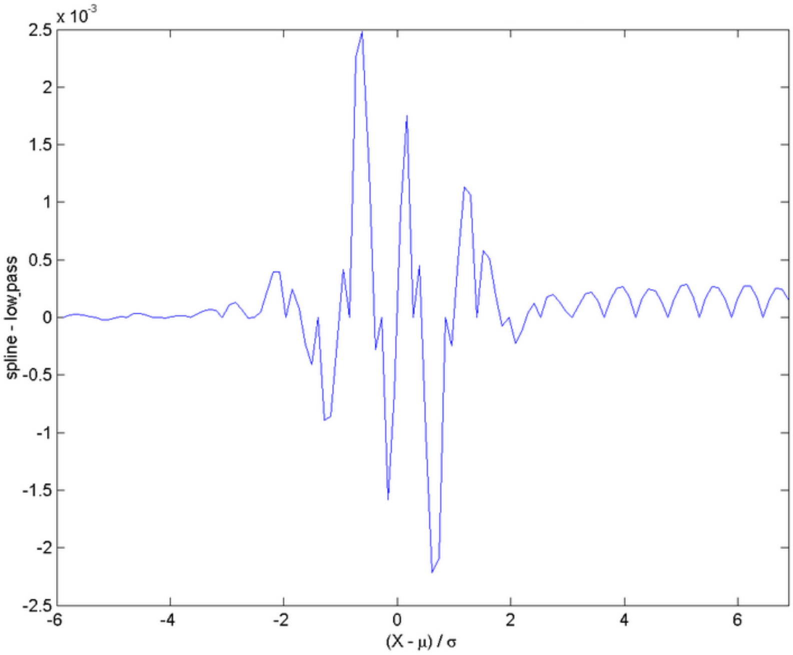
The coefficients used for the *spline* fitting curve are shown in table 5.3.

[1] Values $X = -10$ and $X = 10$ are considered to be high enough compared to the maximal values and have been used to impose the boundary conditions

Table 5.3: coefficients for the polynomials of the spline approximation

	$f(x) = a_0x^3 + a_1x^2 + a_2x + a_3$				
x_{i0}	a_0	a_1	a_2	a_3	x_{if}
-7.00	-0.0003	0.0003	0.0000	0.0000	-5.88
-5.88	0.0020	-0.0006	-0.0003	0.0000	-5.32
-5.32	-0.0043	0.0028	0.0010	0.0000	-4.76
-4.76	0.0073	-0.0043	0.0001	0.0007	-4.20
-4.20	-0.0056	0.0079	0.0022	0.0007	-3.64
-3.64	0.0018	-0.0016	0.0057	0.0034	-3.08
-3.08	0.0195	0.0015	0.0057	0.0064	-2.52
-2.52	-0.0338	0.0343	0.0258	0.0135	-1.96
-1.96	0.1079	-0.0224	0.0325	0.0328	-1.40
-1.40	-0.1166	0.1588	0.1088	0.0629	-0.84
-0.84	0.4241	-0.0371	0.1770	0.1531	-0.28
-0.28	-0.8040	0.6754	0.5344	0.3151	0.28
0.28	0.4241	-0.6754	0.5344	0.6849	0.84
0.84	-0.1166	0.0371	0.1770	0.8469	1.40
1.40	0.1079	-0.1588	0.1088	0.9371	1.96
1.96	-0.0338	0.0224	0.0325	0.9672	2.52
2.52	0.0195	-0.0343	0.0258	0.9865	3.08
3.08	0.0018	-0.0015	0.0057	0.9936	3.64
3.64	-0.0056	0.0016	0.0057	0.9966	4.20
4.20	0.0073	-0.0079	0.0022	0.9993	4.76
4.76	-0.0043	0.0043	0.0001	0.9993	5.32
5.32	0.0022	-0.0029	0.0009	1.0000	5.88
5.88	-0.0006	0.0008	- 0.0002	1.0000	6.44
6.44	0.0002	-0.0002	0.0001	1.0000	7.00

Figure 5.6: Difference between both fitting methods


x 10^-3
spline - low_pass
(X - μ) / σ