Statistical Analysis of Processing Data for a Manufacturing Industry (A Case Study of Stephens Bread Industry)

[1]Ezeliora Chukwuemeka Daniel, [2]Iwenofu Chinwe Onyedika, [3]Offor Ikechukwu Christian and [4]Udoye Benjamin Onyebuchi

[1]Industrial and Production Engineering Department, Nnamdi Azikiwe University Awka, Anambra State
[2,3,4]Mechanical Engineering Department, Federal Polytechnic Oko, Anambra State
[1]cezeliora@gmail.com

***Abstract*– The research work was based on the statistical analysis of the data using statistical tools. The use of nonparametric test was used to extract information from the data. However, the reliability test was used to show how reliable the data is for analysis or modeling. Furthermore, the missing value tests were also used to observe whether the data is a complete or incomplete data. The essence of the statistical analysis is to communicate with the data and to understand the situations that this data portray.**

***Keywords*– Giant Loaf, Long loaf, Small Loaf, Nonparametric Test, Chi-square, P-value, Statistics, Reliability Test and Missing Value Test**

I. INTRODUCTION TO STATISTICS

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data [1], [2]. It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments [1].

The word statistics, when referring to the scientific discipline, is singular, as in "Statistics is an art" [3]. This should not be confused with the word statistic, referring to a quantity (such as mean or median) calculated from a set of data, [4] whose plural is statistics ("this statistic seems wrong" or "these statistics are misleading").

The objective of statistical analysis is to extract information from data in order to better understand the situations that these data portray**.**

Overview of Statistics: In applying statistics to a scientific, industrial, or societal problem, it is necessary to begin with a population or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". A population can also be composed of observations of a process at various times, with the data from each observation serving as a different member of the overall group. Data collected about this kind of "population" constitutes what is called a time series.

For practical reasons, a chosen subset of the population called a sample is studied—as opposed to compiling data about the entire group (an operation called census). Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting. This data can then be subjected to statistical analysis, serving two related purposes: description and inference.

- Descriptive statistics summarize the population data by describing what was observed in the sample numerically or graphically. Numerical descriptors include mean and standard deviation for continuous data types, while frequency and percentage are more useful in terms of describing categorical data.
- Inferential statistics uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of: answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation) and modeling relationships within the data. Inference can extend to forecasting, prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data, and can also include data mining [5].

The concept of correlation is particularly noteworthy for the potential confusion it can cause. Statistical analysis of a data set often reveals that two variables (properties) of the population under consideration tend to vary together, as if they were connected. For example, a study of annual income that also looks at age of death might find that poor people tend to have shorter lives than affluent people. The two variables are said to be correlated; however, they may or may not be the cause of one another. The correlation phenomena could be caused by a third, previously unconsidered phenomenon, called a lurking variable or confounding variable. For this reason, there is no way to immediately infer the existence of a causal relationship between the two variables. To use a sample as a guide to an entire population, it is important that it truly represent the overall population. Representative sampling assures that inferences and conclusions can safely extend from the sample to the population as a whole. A major problem lies in determining the extent that the sample chosen is actually representative.

Statistics offers methods to estimate and correct for any random trending within the sample and data collection procedures. There are also methods of experimental design for experiments that can lessen these issues at the outset of a study, strengthening its capability to discern truths about the population.

Randomness is studied using the mathematical discipline of probability theory. Probability is used in "mathematical statistics" (alternatively, "statistical theory") to study the sampling distributions of sample statistics and, more generally, the properties of statistical procedures. The use of any statistical method is valid when the system or population under consideration satisfies the assumptions of the method.

Misuse of statistics can produce subtle, but serious errors in description and interpretation—subtle in the sense that even experienced professionals make such errors, and serious in the sense that they can lead to devastating decision errors. For instance, social policy, medical practice, and the reliability of structures like bridges all rely on the proper use of statistics. See below for further discussion.

Even when statistical techniques are correctly applied, the results can be difficult to interpret for those lacking expertise. The statistical significance of a trend in the data—which measures the extent to which a trend could be caused by random variation in the sample—may or may not agree with an intuitive sense of its significance. The set of basic statistical skills that people need to deal with information in their everyday lives properly is referred to as statistical literacy.

p-value: In statistical significance testing the **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true [6]. One often "rejects the null hypothesis" when the p-value is less than the predetermined significance level which is often 0.05 [7], [8] or 0.01, indicating that the observed result would be highly unlikely under the null hypothesis (i.e., the observation is highly unlikely to be the result of random chance). Many common statistical tests, such as chi-squared tests or Student's t-test, produce test statistics which can be interpreted using p-values.

The p-value is a key concept in the approach of Ronald Fisher, where he uses it to measure the weight of the data against a specified hypothesis, and as a guideline to ignore data that does not reach a specified significance level. Fisher's approach does not involve any alternative hypothesis, which is instead the Neyman–Pearson approach. The p-value should not be confused with the Type I error rate (false positive rate) α in the Neyman–Pearson approach – though α is also called a "significance level" and is often 0.05, these terms have different meanings, these are incompatible approaches, and the numbers p and α cannot meaningfully be compared. There is a great deal of confusion and misunderstanding on this point, and many misinterpretations, discussed below [9]. Fundamentally, the p-value does not in itself allow reasoning about the probabilities of hypotheses (this requires a prior, as in Bayesian statistics), nor choosing between different hypotheses (this is instead done in Neyman–Pearson statistical hypothesis testing) – it is simply a measure of how likely the data is to have occurred by chance, assuming the null hypothesis is true.

Despite the above caveats, statistical hypothesis tests making use of p-values are commonly used in many fields of science and social sciences, such as economics, psychology, [10] biology, criminal justice and criminology, and sociology, [11] though this is criticized .

Definition of p-value: In brief, the (left-tailed) p-value is the quantile of the value of the test statistic, with respect to the sampling distribution under the null hypothesis. The right-tailed p-value is one minus the quantile, while the two-tailed p-value is twice whichever of these is smaller. This is elaborated below:

Computing a p-value requires a null hypothesis, a test statistic (together with deciding if one is doing one-tailed test or a two-tailed test), and data. The key preparatory computation is computing the cumulative distribution function (CDF) of the sampling distribution of the test statistic under the null hypothesis; this may depend on parameters in the null distribution and the number of samples in the data. The test statistic is then computed for the actual data, and then its quantile computed by inputting it into the CDF. This is then normalized as follows:

- one-tailed (left tail): quantile, value of cumulative distribution function (since values close to 0 are extreme);
- one-tailed (right tail): one minus quantile, value of complementary cumulative distribution function (since values close to 1 are extreme: 0.95 becomes 0.05);
- two-tailed: twice p-value of one-tailed, for whichever side value is on (since values close to 0 or 1 are both extreme: 0.05 and 0.95 both have a p-value of 0.10, as one adds the tails on both sides).

Even though computing the test statistic on given data may be easy, computing the sampling distribution under the null hypothesis, and then computing its CDF is often a difficult computation. Today this computation is done using statistical software, often via numeric methods (rather than exact formulas), while in the early and mid 20th century, this was instead done via tables of values, and one interpolated or extrapolated p-values from these discrete values. Rather than using a table of p-values, Fisher instead inverted the CDF, publishing a list of values of the test statistic for given fixed p-values; this corresponds to computing the quantile function (inverse CDF).

Interpretation of p-value: Hypothesis tests, such as Student's t-test, typically produce test statistics whose sampling distributions under the null hypothesis are known. For instance, in the above coin-flipping example, the test statistic is the number of heads produced; this number follows a known binomial distribution if the coin is fair, and so the probability of any particular combination of heads and tails can be computed. To compute a p-value from the test statistic, one must simply sum (or integrate over) the probabilities of more extreme events occurring. For commonly used statistical tests, test statistics and their corresponding p-values are often tabulated in textbooks and reference works.

Traditionally, following Fisher, one rejects the null hypothesis if the p-value is less than or equal to a specified significance level, [1] often 0.05 [12], or more stringent values, such as 0.02 or 0.01. These numbers should not be

confused with the Type I error rate α in Neyman–Pearson-style statistical hypothesis testing; see misunderstandings, below. A significance level of 0.05 would deem extraordinary any result that is within the most extreme 5% of all possible results under the null hypothesis. In this case a p-value less than 0.05 would result in the rejection of the null hypothesis at the 5% (significance) level.

II. METHODOLOGY

The research methods used were the nonparametric test, the reliability test and the missing values analysis.

Table 1: Process Data for Stephens Bread Industry

Size of loaves		Giant loaf (x1)	long loaf (x2)	small loaf (x3)
Process Time (Mins) Per Loaf Size	Mixing (mins)	2	1	2
	Matching (mins)	1	1	1
	Molding (mins)	3	4	2
	Baking (mins)	4	3	2
Profit per loaf (kobo)		1500	1200	500

Source: Field Research

Non Parametric Tests

Table 2: Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
VAR00004	5	302.0000	669.70329	1.00	1500.00
VAR00005	5	241.8000	535.65166	1.00	1200.00
VAR00006	5	101.4000	222.82459	1.00	500.00

Chi-Square Test

Frequencies

Table 3: VAR00004

	Observed N	Expected N	Residual
1.00	1	1.0	.0
2.00	1	1.0	.0
3.00	1	1.0	.0
4.00	1	1.0	.0
1500.00	1	1.0	.0
Total	5		

a. 5 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.0.

Table 4: VAR00005

	Observed N	Expected N	Residual
1.00	2	1.3	.8
3.00	1	1.3	-.3
4.00	1	1.3	-.3
1200.00	1	1.3	-.3
Total	5		

b. 4 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.3.

Table 5: VAR00006

	Observed N	Expected N	Residual
1.00	1	1.7	-.7
2.00	3	1.7	1.3
500.00	1	1.7	-.7
Total	5		

c. 3 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.7.

Table 6: Test Statistics

	VAR00004	VAR00005	VAR00006
Chi-Square	.000[a]	.600[b]	1.600[c]
df	4	3	2
Asymp. Sig.	1.000	.896	.449

Data Reliability Test

Scale: ALL VARIABLES

Table 7: Case Processing Summary

		N	%
Cases	Valid	5	100.0
	Excluded[a]	0	.0
	Total	5	100.0

a. Listwise deletion based on all variables in the procedure.

Table 8: Reliability Statistics

Cronbach's Alpha	N of Items
.923	3

Automated Data Preparation

Table 9: Case Processing Summary

	N	Percent
Included	5	100.0%
Excluded	0	0.0%
Total	5	100.0%

Analysis of Missing Values

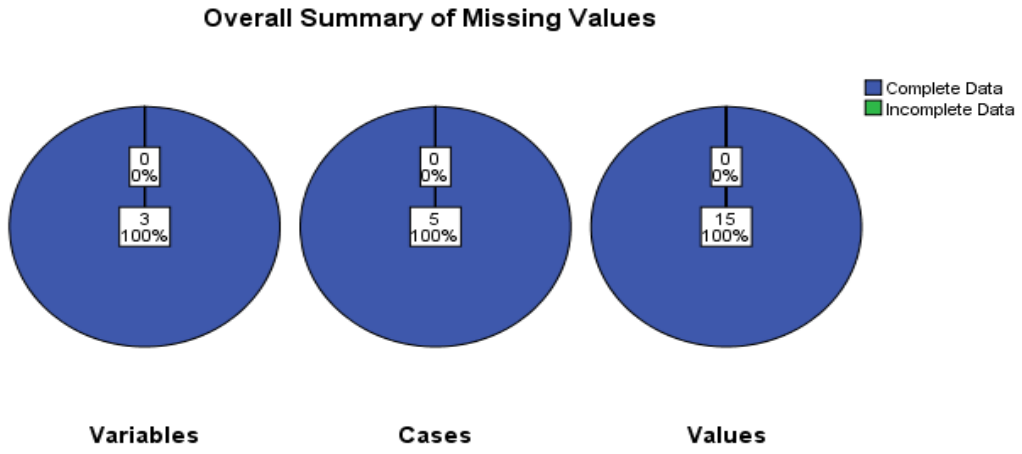

Fig. 1: Analysis of Missing Values

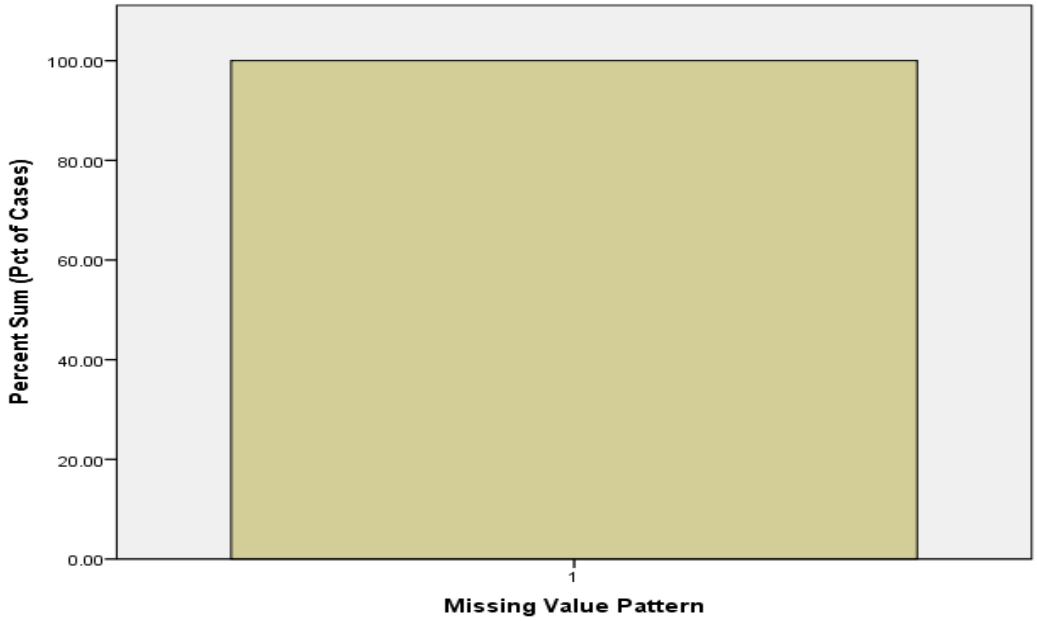

Fig. 2: Percentage sum of the cases of complete data

Descriptive Statistics: Giant loaf (x1), long loaf (x2), small loaf (x3)

Variable	N	N*	CumPct	Mean	SE Mean	StDev	Variance	CoefVar
Giant loaf (x1)	5	0	100	302	300	670	448503	221.76
long loaf (x2)	5	0	100	242	240	536	286923	221.53
small loaf (x3)	5	0	100	101.4	99.7	222.8	49650.8	219.75

Variable	Sum	Sum of Squares	Minimum	Q1	Median	Q3	Maximum
Giant loaf (x1)	1510	2250030	1	2	3	752	1500
long loaf (x2)	1209	1440027	1	1	3	602	1200
small loaf (x3)	507.0	250013.0	1.0	1.5	2.0	251.0	500.0

Variable	Skewness	Kurtosis	MSSD
Giant loaf (x1)	2.24	5.00	279753
long loaf (x2)	2.24	5.00	179102
small loaf (x3)	2.24	5.00	31000.8

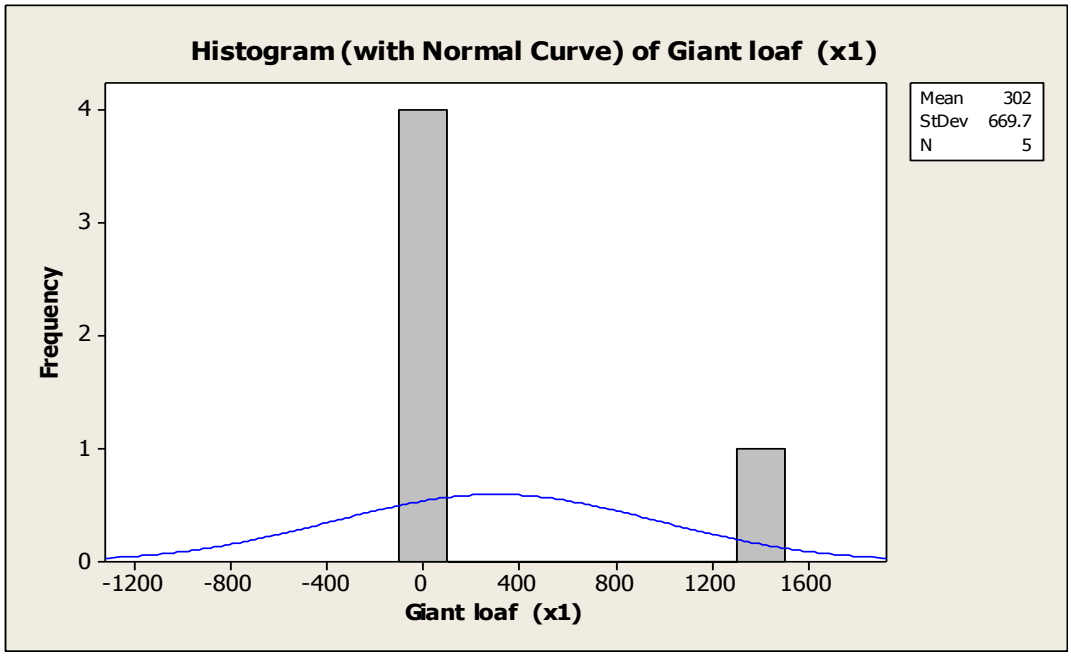

Fig. 3: Histogram (with Normal Curve) of Giant loaf (x1)

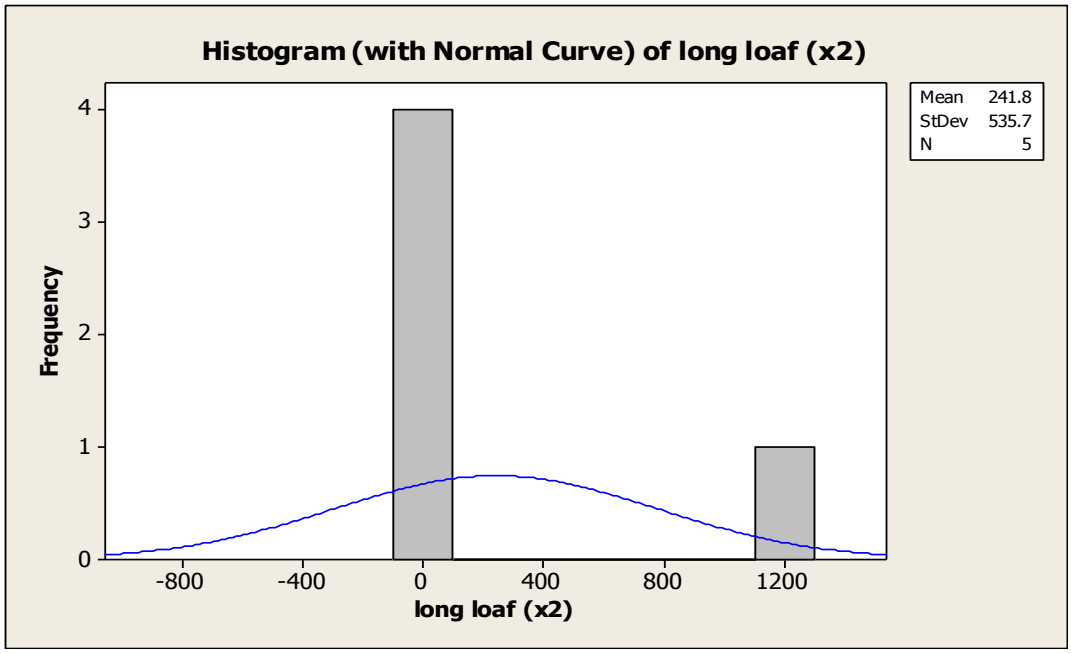

Fig. 4: Histogram (with Normal Curve) of long loaf (x2)

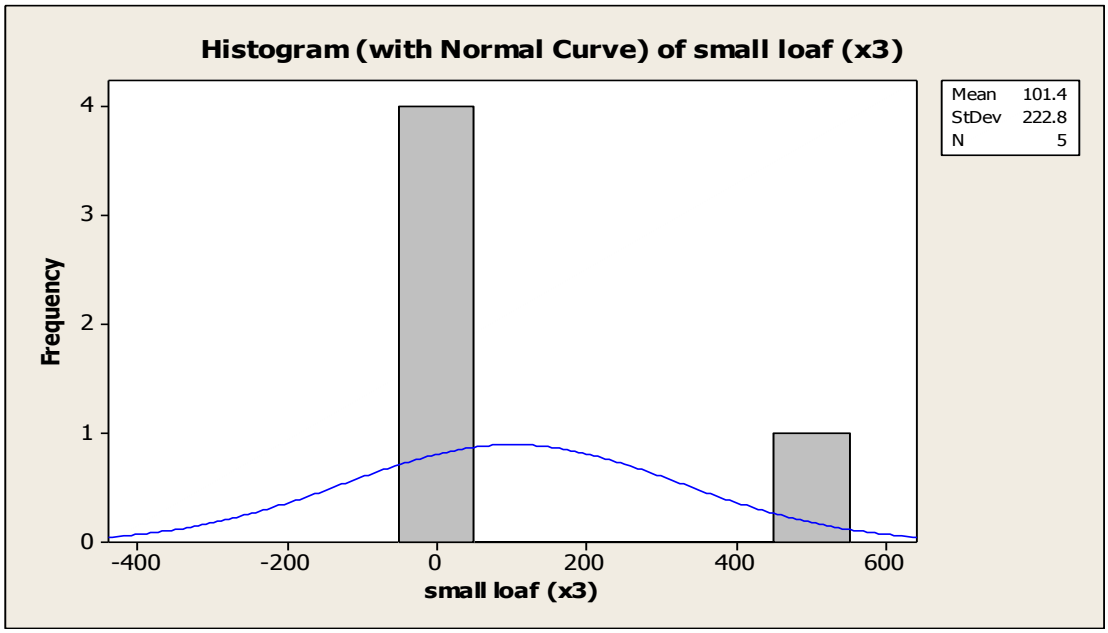

Fig. 5: Histogram (with Normal Curve) of small loaf (x3)

III. DISCUSSION OF RESULTS

From the results, it was observed that the non-parametric tests were used to show the descriptive statistical of the data, while the chi-square shows the observed and the expected values of the data. However, the reliability tests were used to experiment how reliable the collected data for the analysis are. From the reliability test results, it shows that the data is hundred percent (100%) reliable for modeling and analysis. Furthermore, the missing value tests were used to confirm and to checkmate if there is any missing value, to confirm the level of the missing data and to ensure that none of the value(s) (or data) required for the analysis or modeling is/are not missing.

IV. CONCLUSION

Statistical analysis of any data is a key to understand and to communicate with the data. The data analyzed in this research work shows the descriptive statistical analysis of the data, reliability analysis of the data and missing values analysis of the data. These tests help us to understand and to communicate with the data. It is therefore highly recommended to test data statistically before any other further analysis or modeling of the data.

REFERENCES

[1]. Goodman, SN (1999). "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy", Annals of Internal Medicine 130: 995–1004.
[2]. Dallal 2012, Note 31: Why P=0.05?
[3]. Wetzels, R.; Matzke, D.; Lee, M. D.; Rouder, J. N.; Iverson, G. J.; Wagenmakers, E. -J. (2011). "Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests". Perspectives on Psychological Science 6 (3): 291. doi:10.1177/1745691611406923. edit
[4]. Babbie, E. (2007). The practice of social research 11th ed. Thomson Wadsworth: Belmont, CA.
[5]. Fisher 1925, pp. 78–79, 98, Chapter IV. Tests of Goodness of Fit, Independence and Homogeneity; with Table of χ^2, Table III. Table of χ^2.
[6]. Fisher 1971, II. The Principles of Experimentation, Illustrated by a Psycho-physical Experiment.
[7]. Fisher 1971, Section 7. The Test of Significance.
[8]. Fisher 1971, Section 12.1 Scientific Inference and Acceptance Procedures.
[9]. Sterne, J. A. C.; Smith, G. Davey (2001). "Sifting the evidence–what's wrong with significance tests?". BMJ (Clinical research ed.) 322 (7280): 226–231. doi:10.1136/bmj.322.7280.226. PMC 1119478. PMID 11159626. edit
[10]. Schervish, M. J. (1996). "P Values: What They Are and What They Are Not". The American Statistician 50 (3). doi:10.2307/2684655. JSTOR 2684655. edit
[11]. Casella, George; Berger, Roger L. (1987). "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem". Journal of the American Statistical Association 82 (397): 106–111.
[12]. Sellke, Thomas; Bayarri, M. J.; Berger, James O. (2001). "Calibration of p Values for Testing Precise Null Hypotheses". The American Statistician 55 (1): 62–71. doi:10.1198/000313001300339950. JSTOR 2685531. edit