目錄

[Part1\_step 2](#_Toc523154250)

[Part2\_sas\_program 5](#_Toc523154251)

[part3\_index1\_abnormal 7](#_Toc523154252)

[Part3\_index2\_beta 11](#_Toc523154253)

[Part3\_index3\_distribution 11](#_Toc523154254)

[Part3\_Index4\_nonparametric 23](#_Toc523154255)

[Part3\_index5\_time series 24](#_Toc523154256)

[Part4\_Index6\_cluster and classfication 28](#_Toc523154257)

**Part1 step**

步驟

1. 是否符合正態分佈

通過適當的變形而符合正態分佈類型：

1. 不能直接以正態分佈顯示的原因（index1），當中的具體原因的含義，即所選的資料特點包含的各種維度，瞭解充分的維度有哪些，從而進一步瞭解哪些資料特點是正態分佈的。
2. 非正態分佈中的所有其他分佈是否有聯繫，大多通過beta函數 exponential函數聯繫著，證明他們之間同屬一個性質，只在不同的層面上較有特點的顯示出來（index2）。
3. 從所得的維度和資料特徵能夠選出資料的合適分佈方式
4. 在樣本無確定分佈下可使用非參數統計方法

獲取每一個正態分佈資料所擁有的二項式本質，中心極限定理認定資料量的增大使得資料的概率密度分佈趨於正態分佈，因此，用二項式本質的角度得出極限定理的意思解釋

z檢驗為常態檢驗，使用總體的方差和均值計算（根據中心極限定理得出在大樣本下，均值和方差逼近總體正式均值方差）,此值服從正態分佈，在小樣本或者總體方差未知下，使用樣本方差計算等式Z值服從T分佈，使用T分佈單樣本t檢驗（sas\_index）

從正態分佈延伸至卡方分佈、T分佈、F分佈（index3）

1. 不同變數之間的聯繫及檢驗分佈

假設檢驗的方法能夠看出變數間的差異或者聯繫（sas\_index）

1. 分類與分類之間
2. 卡方檢驗
3. 分類與連續之間
4. 使用非參數統計方法（Nonparametric statistics）（index4）
5. 決策樹方法（連續變數可分段為離散變數，方差分析）
6. 連續與連續之間
7. 多變數正態化後的聯合概率分佈與相關性
8. 單元和多元的線性回歸方法
9. 主成分分析
10. 多元回歸
11. 多元線性回歸（index3）
12. 邏輯回歸（index6）
13. 时间序列（index5）
14. 分類聚類

非參數的方法，傅裡葉擬合方法（kernal regression）

k-mean聚類

神經網絡

1. 有效的差異形成不同群體

EM算法

**Part2 Index**

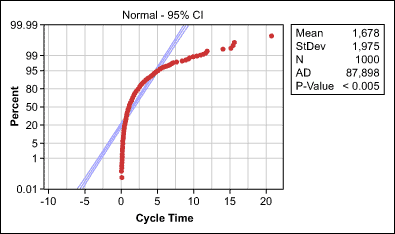
**Index-1： Abnormal Distribution**

Normally distributed data is a commonly misunderstood concept in Six Sigma. Some people believe that all data collected and used for analysis must be distributed normally. But normal distribution does not happen as often as people think, and it is not a main objective. Normal distribution is a means to an end, not the end itself.

Normally distributed data is needed to use a number of statistical tools, such as individuals control charts, Cp/Cpk analysis, t-tests and the analysis of variance (ANOVA). If a practitioner is not using such a specific tool, however, it is not important whether data is distributed normally. The distribution becomes an issue only when practitioners reach a point in a project where they want to use a statistical tool that requires normally distributed data and they do not have it.

The probability plot in Figure 1 is an example of this type of scenario. In this case, normality clearly cannot be assumed; the p-value is less than 0.05 and more than 5 percent of the data points are outside the 95 percent confidence interval.

Figure 1: Probability Plot of Cycle Time



What can be done? Basically, there are two options:

Identify and, if possible, address reasons for non-normality or

Use tools that do not require normality

Addressing Reasons for Non-normality

When data is not normally distributed, the cause for non-normality should be determined and appropriate remedial actions should be taken. There are six reasons that are frequently to blame for non-normality.

**Reason 1: Extreme Values**

Too many extreme values in a data set will result in a skewed distribution. Normality of data can be achieved by cleaning the data. This involves determining measurement errors, data-entry errors and outliers, and removing them from the data for valid reasons.

It is important that outliers are identified as truly special causes before they are eliminated. Never forget: The nature of normally distributed data is that a small percentage of extreme values can be expected; not every outlier is caused by a special reason. Extreme values should only be explained and removed from the data if there are more of them than expected under normal conditions.

**Reason 2: Overlap of Two or More Processes**

Data may not be normally distributed because it actually comes from more than one process, operator or shift, or from a process that frequently shifts. If two or more data sets that would be normally distributed on their own are overlapped, data may look bimodal or multimodal – it will have two or more most-frequent values.

The remedial action for these situations is to determine which X’s cause bimodal or multimodal distribution and then stratifythe data. The data should be checked again for normality and afterward the stratified processes can be worked with separately.

An example: The histogram in Figure 2 shows a website’s non-normally distributed load times. After stratifying the load times by weekend versus working day data (Figure 3), both groups are normally distributed.

Figure 2: Website Load Time Data

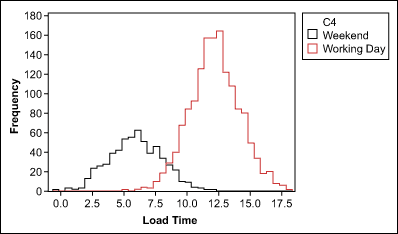
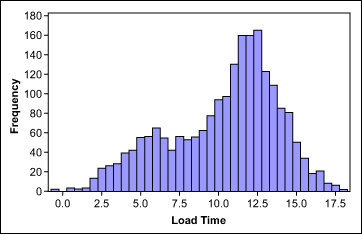


Figure 3: Website Load Time Data After Stratification

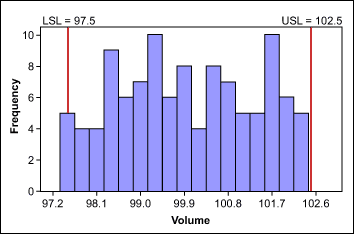
**Reason 3: Insufficient Data Discrimination**

Round-off errors or measurement devices with poor resolution can make truly continuous and normally distributed data look discrete and not normal. Insufficient data discrimination – and therefore an insufficient number of different values – can be overcome by using more accurate measurement systems or by collecting more data.

**Reason 4: Sorted Data**

Collected data might not be normally distributed if it represents simply a subset of the total output a process produced. This can happen if data is collected and analyzed after sorting. The data in Figure 4 resulted from a process where the target was to produce bottles with a volume of 100 ml. The lower and upper [specifications](http://www.isixsigma.com/dictionary/specification) were 97.5 ml and 102.5 ml. Because all bottles outside of the specifications were already removed from the process, the data is not normally distributed – even if the original data would have been.

Figure 4: Sorted Bottle Volume Data



**Reason 5: Values Close to Zero or a Natural Limit**

If a process has many values close to zero or a natural limit, the data distribution will skew to the right or left. In this case, a transformation, such as the Box-Cox power transformation, may help make data normal. In this method, all data is raised, or transformed, to a certain exponent, indicated by a Lambda value. When comparing transformed data, everything under comparison must be transformed in the same way.

The figures below illustrate an example of this concept. Figure 5 shows a set of cycle-time data; Figure 6 shows the same data transformed with the natural logarithm.

Figure 5: Cycle Time Data

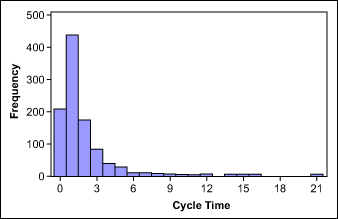
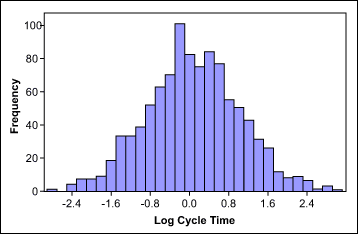


Figure 6: Log Cycle Time Data



Take note: None of the transformation methods provide a guarantee of a normal distribution. Always check with a probability plot to determine whether normal distribution can be assumed after transformation.

**Reason 6: Data Follows a Different Distribution**

There are many data types that follow a non-normal distribution by nature. Examples include:

Weibull distribution, found with life data such as survival times of a product

Log-normal distribution, found with length data such as heights

Largest-extreme-value distribution, found with data such as the longest down-time each day

Exponential distribution, found with growth data such as bacterial growth

Poisson distribution, found with rare events such as number of accidents

Binomial distribution, found with “proportion” data such as percent defectives

If data follows one of these different distributions, it must be dealt with using the same tools as with data that cannot be “made” normal.

**Index-2：Normal Distribution**

**Binomial Distribution**

**By De Moivre–Laplace Theorem**

**Descriptive**

Sample:

Population:

**Algebra of random variables-noraml distribution**

Base：

By Moment-generating function:

By Moment-generating function:

Algebra of independent:

By Characteristic Function:

By Law of total expectation:

Combination：

**Chi-squared Distribution**

By Characteristic Function:

**Student's t-distribution**

**F-distribution**

**Linear regression（condition distribution）**

Simple linear regression

By maxlikelihood estimate

By Linear least squares

By Gauss–Markov theorem

Estimate of residual variance by projection

By ols

And by

so

The F-test for Linear Regression

Sum of squares for regression

Sum of squares for residuals

Sum of squares total

**Index-3：Beta Distribution**

**Index-4：Distribution Formula Derivation**

**De Moivre–Laplace Theorem**

**Stirling's approximation**

Conclusion：

Proof：

By

By

**Gaussian Integral**

**Euler's number**

**MacLaurin series**

**Dependent and independent variables**

Independent:

When:

So:

Dependent(regress):

**Conditional probability distribution**

**Law of total expectation**

**Characteristic Function**

Proof：

**Moment-generating function**

Proof:

Chi-squared Distribution Characteristic Function

Proof：

**Convolution theorem**

**So:**

**Cochran's theorem**

**Gauss–Markov theorem**

**Linear least squares**

**Projection on OLS**

**Index5 Nonparametric statistics**

Wilcoxon Signed Rank

boxcox

* Box GEP, Cox DR. [An analysis of transformations.](http://www.jstor.org/stable/2984418) J R Stat Soc B. 1964; 26(2):211-52.
* Koch AL. [The logarithm in biology. 1. Mechanisms generating the log-normal distribution exactly.](http://dx.doi.org/10.1016/0022-5193%2866%2990119-6) J Theor Biol. 1966; 12(2):276-90.
* Koch AL. [The logarithm in biology. II. Distributions simulating the log-normal.](http://dx.doi.org/10.1016/0022-5193%2869%2990040-X) J Theor Biol. 1969; 23(2):251-68.
* McAlister D. [The law of the geometric mean.](http://www.jstor.org/stable/113784) Proc R Soc London. 1879; 29:367-76.

Johnson Transformation

**Index6 time series**

Box–Jenkins method

Autoregressive model

Moving-average

Arma

Arima

Autocorrelation

Of series

Of MA

Of AR

Partial autocorrelation function

By ols

Yule-Walker Equations

So

Partial autocorrelation function

**Index7 cluster and classfication**

**Principle of maximum Entropy**

**Maximum likelihood estimation**

**Logistic Regression Models**

By Maximum likelihood estimation

Using Gradient descent to find

**Part3 SAS\_Program**

**SAS\_Program**

分類變數 for a

有序變數 for b

Chi-square test：a-a

Student's t-test: a-b

Analysis of variance:a-b

Reg:b-b

**/\*univariate\*/**

**proc** **univariate** data=j normal;

histogram CARD\_SAL\_AVG\_N/ normal midpoints=-**1** to **1** by **0.01**;

**run**;

**/\*mixdistribute\*/**

ods graphics on;

**proc** **fmm** data=WORK.QUERY\_FOR\_INFO\_0003 plots=all;

where clsname="家具";

model logamount= /dist=normal k=**2**;

**run**;

ods graphics off;

**/\*t-test\*/**

ods graphics on;

**proc** **ttest** data=a h0=**3** side=**2** plots=histogram;

var c;

**run**;

ods graphics off;

**proc** **ttest** data=a;

class a;

var b;

**run**;

**/\*chisq\*/**

**proc** **freq** data=a;

table a\*b / chisq;

**run**;

**/\*anova\*/**

ods graphics on;

**proc** **anova** data=a PLOTS(MAXPOINTS=none);

class a;

model b=a;

**run**;

ods graphics off;

**/\*time\_series\*/**

**proc** **arima** data=a.s;

identify var=money(**1**,**1**) nlog=**30** minic p=(**0**:**10**) q=(**0**:**10**);

estimate p=**1** q=**2**;

forecast lead=**31** out=a.results;

**run**;

statement

a,

identify var=x(**1**,**1**)

which is

b,

identify var=x(**2**)

which is

C,

estimate p=**1** q=**2**;