Indian Academy of Sciences, Bengaluru
Indian National Science Academy, New Delhi
The National Academy of Sciences India, Allahabad

SUMMER RESEARCH FELLOWSHIPS — 2017

Format for the final Report*

Name of the candidate : MISS. MRUDULA GAJANAN BELGAL

Application Registration no. : MATS 625

Date of joining : MAY 18th, 2017

Date of completion : JULY 14th, 2017

Total no. of days worked : 58

Name of the guide : DR. INDRANIL MUKHOPADHYAY

Guide's institution : INDIAN STATISTICAL INSTITUTE, KOLKATA

Project title : A STUDY OF ZERO-INFLATED DISTRIBUTIONS

Address with pin code to which the certificate could be sent:

B003, G.M. PATNE SANKUL, KUMBHAR ALI, A/p. KHED, DIST. RATNAGIRI, MAHARASHTRA. PIN - 415709

E-mail ID: mrudulabelgal@gmail.com

Phone No: 9209896839

TA Form attached with final report : YES ______ NO ✓

If, NO, Please specify reason : TRAVELLED WITH PARENTS

MGBelgal
Signature of the candidate
Date: 14th July 2017

Indsathy
Signature of the guide
Date: 14/07/2017

*The final report could be anywhere between 20 and 25 pages including tables, figures etc.
This format should be the first page of the report and should be stapled with the main report.

(For office use only; do not fill/tear)

Candidate's name:	Fellowship amount:
Student: Teacher:	Deduction:
Guide's name:	TA fare:
KVPY Fellow: INSPIRE Fellow:	Amount to be paid:
Others	A/c holder's name:

IMPORTANT NOTES:
A soft copy of this report should be uploaded in the online page of our website by making use of the userid/password provided to you.

INDIAN STATISTICAL INSTITUTE

KOLKATA, INDIA.

A STUDY OF ZERO-INFLATED DISTRIBUTIONS

FINAL REPORT

By

Miss Mrudula Belgal.

Fergusson College, Pune 04.

Application no. MATS625.

ABSTRACT

Under this Summer Research Fellowship I started working in Human Genetics Unit, ISI, Kolkata, as a trainee under Dr. Indranil Mukhopadhyay from May 18th, 2017. In this project, efforts are taken to get an insight of two topics, tests related to Zero Inflated Distributions and Functional Data Analysis. Both the topics have their own importance in near future.

Zero is an important figure in every science that exists, Statistics is not an exception. Zeros can be helpful and troublesome in many ways. One such way is obtaining large number of zeros in observations taken for an experiment. It harms the possibility of accurate analysis and predictions and hence needs to be tackled in the beginning. Thus, they found a way to incorporate the extra zeros in the analysis and then comes the testing of hypotheses to check its truthfulness. I studied the background of zero inflation and inferential statistics related to such situations.

Another part of the study is Functional Data Analysis which has speedy development because of its application in wide range of current problems. I tried to study the nature of functional data and some basic inferential techniques related to it.

The period also involved learning and developing software skills and training to understand and read the research papers. This report gives account of the subjects I have studied in this program.

Part I : ZERO INFLATION

Zero Inflation is a case where large, excess number of zeros appear in data. For such data fitting a probability distribution, studying their characteristics needs some extra work than usual.

Examples Explained

1. Let us consider a situation. A park is placed in a town which has a fishing pond. We consider the persons returning from the park and we ask them many questions along with 'How many fish did you catch?'

When we think about the situation statistically, we can see that 'No. of fish caught in particular time interval' must follow Poisson model, which has support from zero. But for this question we have two cases that generates zeros;

i. The persons who went for fishing, but did not catch any.
ii. The persons who did not go for fishing at all.

The first case gives the zeros from the Poisson model, but the second case generates zeros that are not included in a Poisson fit and hence are excess.

2. When an insurance company considers a population in an area and collects the data of 'No. of claims of particular type they received on a day', kike above examples zeros are generated by two possibilities;

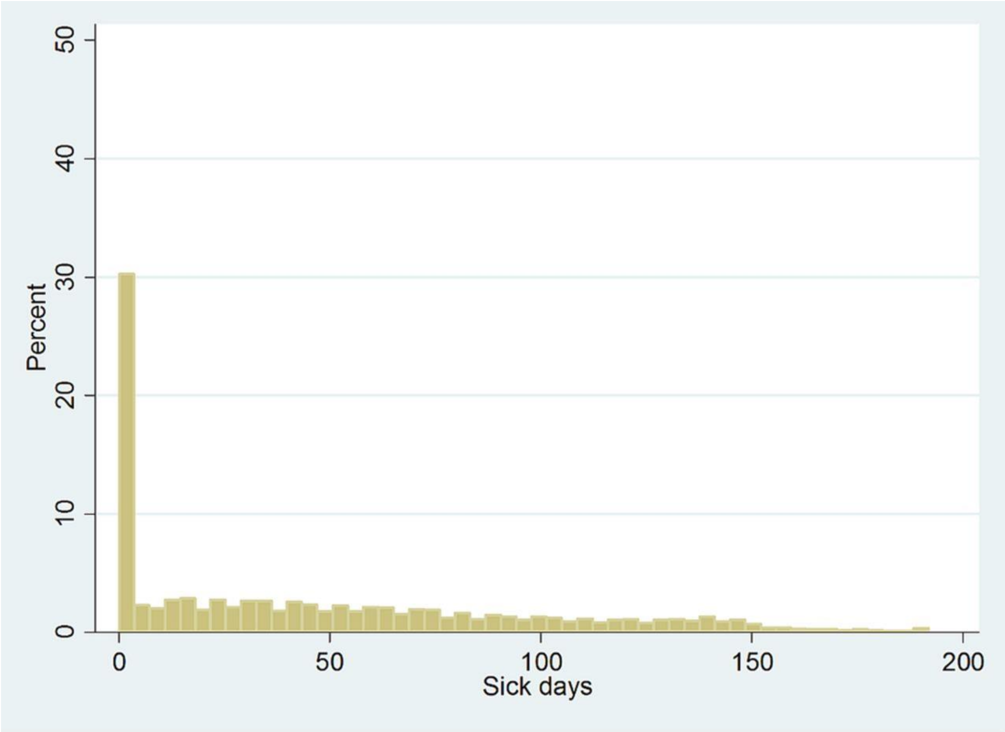
i. From the persons who were insured but didn't claim.
ii. From the persons of the population who didn't have the insurance.

To deal with such a problem, this we use techniques related to Zero Inflation.

3. Let us consider another situation. In a properly designed experiment, the controlled hormonal treatment is given to say a thousand plants, the hormone plays a role in the growth of roots and the number of rootlets grown in 24 hours is observed.

The recorded observations contain plenty of zeros whose number cannot be ignored and hence they must be included in statistical model, which is an unusual case.

4. Let us consider a graph for a situation where we recorded number of days employees of a branch of a company were sick. The graph shows the existence of zero inflation.



TYPES OF ZERO INFLATED DISTRIBUTIONS

Terminology

The population consists of two types of zeros

i. True zeros- Zeros generated by a distribution.
ii. Excess zeros- Zeros other than the zeros from distribution.

For discrete data the distributions may be Poisson Distribution or Negative Binomial Distribution.

Zero Inflated Poisson Model

Let Y be zero inflated Poisson variable.

w be the probability of all the zeros observed.

p be the mean of Poisson count.

The p.m.f. of Y is given by

$$P[Y = y] = w + (1 - w)\, e^{-p} \qquad ; y = 0$$
$$= \frac{(1-w)\, e^{-p}\, p^{y}}{y\,!} \qquad ; y > 0$$
$$= 0 \qquad ; \text{otherwise}$$

We denote it as Y ~ ZIP(p, w)

For Y,

E[Y] = (1 - w) p

Var[Y] = (1 - w) (1 + pw) p

Zero Inflated Negative Binomial Model

Let Y be zero inflated Negative binomial variable.

w be the probability of all the zeros observed.

k, p be the parameters of Negative binomial model.

The p.m.f. of Y is given by

$$
\begin{aligned}
P[Y = y] &= w + (1 - w)(1 + k.p)^{-1/k} && ; y = 0 \\
&= (1 - w)\,\frac{(k\,p)^{y}\,\Gamma(y+1/k)}{\Gamma(y+1)\,\Gamma\left(\frac{1}{k}\right)\,(1+k\,p)^{y+1/k}} && ; y > 0 \\
&= 0 && ; \text{otherwise}
\end{aligned}
$$

We denote it as Y ~ ZINB(k, p, w)

For Y,

E[Y] = (1 - w) p

Var[Y] = (1 - w) (1 – p(w + k)) p

Notes

- From above distributions, three cases related to 'w' are observed;
  - w > 0 – Over-deposition
  - w = 0 – Standard distribution
  - w < 0 – Under-deposition, is a valid probability distribution.
- Some other discrete distributions like Binomial and some continuous distributions whose support aids the occurrence of excess zeros like Exponential distribution follow zero inflated model, and its general set-up is;

  X ~ f(x; θ) ; w is the probability of zeros,

  then,
  $$
  \begin{aligned}
  P[X = x] &= w + (1 - w)\, f(x = 0) && ; x = 0 \\
  &= (1 - w)\, f(x) && ; x > 0 \\
  &= 0 && ; \text{otherwise}
  \end{aligned}
  $$

SOME TESTS RELATED TO ZERO INFLATED POISSON MODEL

1. <u>Test for Over-deposition or Under-deposition</u>

   We know, the mean and variance of random variable X, X~ ZIP (λ, p) are by

$E[X] = (1 - p)\lambda$

$Var[X] = (1 - p)(1 + \lambda p)\lambda$

We are to construct LRT for the hypothesis under consideration

H_0: Data is from Poisson distribution Against

H_1: Data is from ZIP

To construct LRT, we need to estimate λ and p, but no closed form is obtained.

Thus, we construct two the non-parametric test statistics.

Suppose, $X_1, X_2, \ldots, X_n$ is a random sample with sample mean E[X] and sample mean square S^2 .

By asymptotic theory, $\frac{\sqrt{n}\,(S^2 - E[X])}{\sqrt{2}\,\lambda}$ follows N(0, 1)

But under H_0, both E[X] and S_n^2 are consistent for λ and hence, λ in above equation can be replaced by each of them.

Therefore, under H_0, $T_1 = \frac{\sqrt{n}\,(S^2 - E[X])}{(\sqrt{2}\,E[X])}$

$$T_2 = \frac{\sqrt{n}\,(S^2 - E[X])}{(\sqrt{2}\,Sn2)}$$

Using exact variance of (S_n^2 - E[X]), test statistic proposed by Böhning is obtained by replacing λ in expression $\frac{(S^2 - E[X])}{\sqrt{2\lambda^2/(n-1)}}$

The test is also called Neyman Scott test.

2. <u>Another LRT and Wald test</u>

Let us consider a random variable Y, Y ~ ZIP (λ, w).

For observations y_i, i = 1,2,….,n

Log likelihood function is given by

$$l(\lambda, w, y) = \sum_{i=1}^{n} \{ I(y_i = 0) \ln(w_i + (1- w_i) \exp(\lambda_i)) + I(y_i > 0) [\ln(1- w_i) - \lambda_i + y_i \ln \lambda_i - \ln y_i!] \}$$

where, I(.) is an indicator function i.e. it equals to 1 if the event is true and equals to 0 otherwise.

To apply Zero inflated Poisson practically, Lambert suggested a joint models for λ and w

$\ln (\lambda) = X\beta$ and $\ln(\frac{w}{(1-w)}) = G\gamma$

where, X, G – covariate matrices

β, γ – p*1 and q*1 matrices of unknown parameters

Testing if a Poisson model is adequate corresponds to testing

H_0: w = 0 Against

H_1: w > 0

To evaluate test statistics, we need model under alternative hypothesis to be estimated. Hence we estimate λ and w as λ_{est} , w_{est}

For LRT,

$R_w = -2 [l(\lambda_{est}) – l(\lambda_{est} , w_{est})]$

where, $l(\lambda_{est})$ – maximum log likelihood under Poisson distribution

$l(\lambda_{est} , w_{est})$ – maximum log likelihood under Zero inflated Poisson distribution

Under H_0 , R_w follows χ^2 distribution with 'q' degrees of freedom.

For Wald test,

The test statistic is given as

$W_w = w_{est}^T \{ \mathrm{cov} (w_{est}) \}^{-1} w_{est}$

Which in case of single constant parameter w, simplifies to

$$W_w = \frac{w_{est}}{\text{var}\,(w_{est})}$$

Under H_0 , W_w follows χ^2 distribution with 'q' degrees of freedom.

Notes

- Replacing the model alternative $H_1 : w \neq 0$ would give similar standard results for both test statistics.
- It is found that appropriate reference distribution for both R_w and W_w is a mixture of chi- squared distributions.

3. <u>Score test</u>

The more general score vector using the notations mentioned above is given by

$$S(\beta, \lambda) = \begin{bmatrix} S(\beta, \lambda) \\ S(\beta, \lambda) \end{bmatrix} = \begin{bmatrix} \frac{\partial l(\lambda, w)}{\partial \beta} \\ \frac{\partial l(\lambda, w)}{\partial \gamma} \end{bmatrix}$$

where,

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_j} = \sum_{i=1}^{n} \{ I(y_i = 0) \left[-\frac{(1-w_i)e^{-\lambda_i}}{w_i + (1-w_i)e^{-\lambda_i}} \right] \lambda_i + I(y_i > 0)(y_i - \lambda_i) \} x_{ij}$$

j = 1,2,….,p

$$\frac{\partial l}{\partial \gamma_r} = \frac{\partial l}{\partial w_i} \frac{\partial w_i}{\partial \gamma_r} = \sum_{i=1}^{n} \{ I(y_i = 0) \left[-\frac{(1-e^{-\lambda_i})}{w_i + (1-w_i)e^{-\lambda_i}} \right] \lambda_i + I(y_i > 0) \left[-\frac{1}{(1-w_i)} \right] \} g_{ir}$$

r = 1,2,….,q

The expected information matrix *I(β , γ)* is then calculated using partition and is used while calculating the test statistic.

Under the null hypothesis, general score test for ZIP model with constant w is then

$$\frac{\left[\sum_{i=1}^{n} (I(y_i = 0) - e^{-\lambda_{0i}})/e^{-\lambda_{0i}}\right]^2}{\left[\sum_{i=1}^{n} (1 - e^{-\lambda_{0i}})/e^{-\lambda_{0i}}\right] - \lambda_0{}^T X \left[X^T diag(\lambda_0) X\right]^{-1} X^T \lambda_0}$$

where, λ_0 is replaced by $\lambda_{0\text{est}}$ i.e. estimate of it.

In this case, the score test statistic simply compares the observed zero frequency with the expected value under the Poisson model along with appropriate weights.

HURDLE MODEL AND ZERO INFLATED MODEL

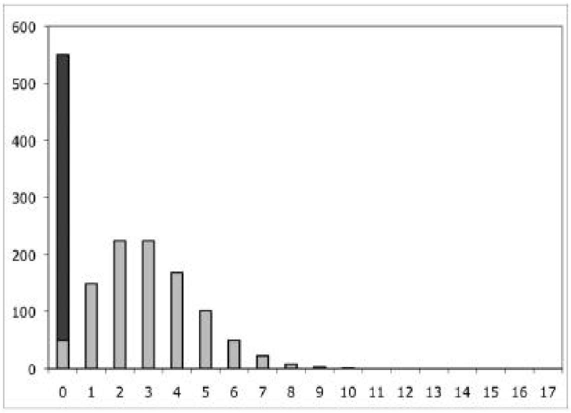
Zero inflated distributions and Hurdle model both serve the purpose of tackling extra zeros in different conceptual ways, so, while studying zero inflation, one cannot neglect the possibility of usefulness of Hurdle model.

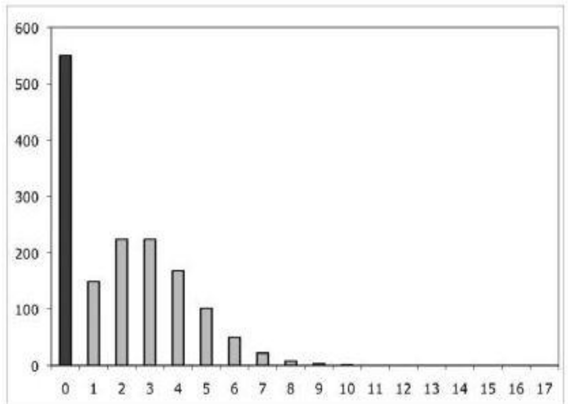
For zero inflated models, the zeros obtained are of two types, structural and sampled, but in hurdle model all zeros are structural in nature. Consider an example, in a study of health organisation, number of cigarettes smoked during a day is recorded for a day before.

For zero inflated situation, observation of zeros will be obtained by non-smokers (structural) and the smokers who did not smoke at all during the day (sampled).

But, in case of hurdle distribution, if the subject is considered to be smoker, then they are not able to score zeros at all and will have positive count, the only source of zeros will be non-smokers (structural), and the hurdle model will incorporate appropriate truncated distributions.

The difference can be shown by graph is





The dark coloured part shows structural zeros and grey part shows sampled zeros.

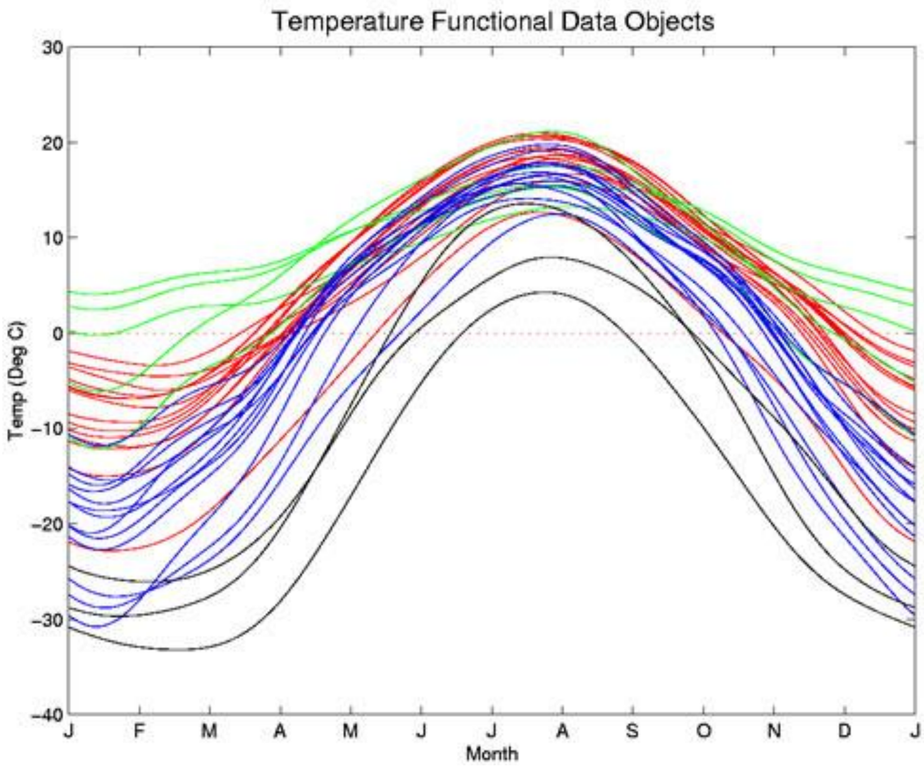
Part II : FUNCTIONAL DATA ANALYSIS

In many experiments, the data obtained may be in the form of curves. Whatever the data is, it is essential to analyse it for future predictions and planning. This introduces a new branch called Functional Data Analysis (FDA).

It is widely applied in many fields such as medical research, biometrics, chemometrics, econometrics etc. in the forms of growth curves, hormone changes and so on. The FDA concerns data in which observations are real functions.

Example Explained

Consider data collected by meteorological department, say of temperature, at different stations in a year. The plot of the data will look like



We can observe different amplitudes, different ranges and different variations corresponding to different places.We consider each observation series $x_i(t)$ and approximate it using the same functional family.

We represent the function x by a linear expansion

$x(t) = \sum_{k=1}^{K} c_k \, \theta_k(t)$ in terms of K known basic functions $\theta_k(t)$.

here, c_i's are coefficients to be estimated using n observations in each series.

SOME TESTS UNDER FUNCTIONAL DATA ANALYSIS

<u>Prerequisites</u>

In functional data analysis, a function data set or a curve from certain stochastic process can be modelled as

$x_i(t) = \mu(t) + \alpha_i(t) + \varepsilon_i(t)$; $i = 1,2,\ldots.,n$

where, x_i's are independent

$\alpha_i(t)$'s is the ith individual function variation from $\mu(t)$, and

$\varepsilon_i(t)$'s is the ith measurement error process or the noise process.

$\alpha(t) \sim SP (0 , \gamma)$

$\varepsilon(t) \sim SP (0 , \gamma_\varepsilon)$, SP denotes stochastic process with mean function $\mu(t)$ and covariance function $\gamma(s, t)$; $s, t \in T$

With no loss of generality, it is assumed that the observed process are not noise process and hence model becomes

$f_i(t) = \mu(t) + \alpha_i(t)$; $i = 1,2,\ldots.,n$

where, f_i's are i.i.d. from underlying stochastic process

i.e. $f(t) \sim SP(\mu, \lambda)$

We are interested in testing the hypotheses about the mean function as they give important statistical inferences.

Assume,

$f(t) = \mu_1(t) + u(t) \sim SP(\mu_1, \lambda_1)$; $f_1(t), f_2(t),\ldots., f_{n1}(t)$ is random sample drawn from it,

$g(t) = \mu_2(t) + v(t) \sim SP(\mu_2, \lambda_2)$; $g_1(t), g_2(t),\ldots., g_{n2}(t)$ is random sample drawn from it,

where, $n_1 + n_2 = n$

We are to test,

H_0: $\mu_1(t) = \mu_2(t)$ for ant $t \in T$ Against

H_1: $\mu_1(t) \neq \mu_2(t)$ for some $t \in T$.

1. Paired t-test

We use pointwise t-test to find whether two groups curves have same mean Functions.

In this method, the estimates are computed pointwise.

Let , for given t, observed curves

$f_1(t), f_2(t), \ldots, f_{n1}(t) \sim AN(\mu_1(t), \sigma_1^2(t))$ are drawn from f(t)

$g_1(t), g_2(t), \ldots, g_{n2}(t) \sim AN(\mu_2(t), \sigma_2^2(t))$ are drawn from g(t).

For any $t \in T$, let

$\mu_1(t)_{\text{estimated}} = n_1^{-1} \sum_{i=1}^{n_1} f_i(t)$

$\mu_1(t)_{\text{estimated}} = n_2^{-1} \sum_{i=1}^{n_2} g_i(t)$

$\sigma_1^2(t)_{\text{estimated}} = (1 - n_1)^{-1} \sum_{i=1}^{n_1} (f_i(t) - \mu_1(t)_{estimate})^2$

$\sigma_2^2(t)_{\text{estimated}} = (1 - n_2)^{-1} \sum_{i=1}^{n_2} (g_i(t) - \mu_2(t)_{estimate})^2$

The test statistic is given by,

$T_n = \frac{\mu_1(t) - \mu_2(t)}{\sqrt{\frac{\sigma_1^2(t)}{n_2} + \frac{\sigma_1^2(t)}{n_2}}}$; all values used are estimated, $\sigma_1^2(t) \neq \sigma_2^2(t)$

$T_n = \frac{\mu_1(t) - \mu_2(t)}{\sqrt{S_p^2\left(\frac{1}{n_2} + \frac{1}{n_2}\right)}}$; all values used are estimated, $\sigma_1^2(t) = \sigma_2^2(t)$

where, $S_p^2 = \frac{[(1-n_1)\sigma_2^2(t)_{estimate} + (1-n_2)\sigma_2^2(t)_{estimate}]}{(n_1 + n_2 - 2)}$

Under null hypothesis, T_n follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom and the rejection of hypothesis depends on level of significance.

Disadvantage of this test is only that the test is for individual time points and overall testing of hypothesis is not possible.

2. <u>L^2- norm based test</u>

Let the observed curves

$f_1(t), f_2(t), \ldots, f_{n1}(t) \sim SP(\mu_1(t), \gamma_1(s,t))$ are drawn from f(t)

$g_1(t), g_2(t), \ldots, g_{n2}(t) \sim SP(\mu_2(t), \gamma_2(s,t))$ are drawn from g(t).

Consider the L^2 norm of difference between $\mu_1(t)_{estimate}$ and $\mu_2(t)_{estimate}$.

The norm will be generally small for valid null hypothesis and large otherwise.

The teat statistic is given by

$T_n = n\,\|\mu_1(t)_{estimate} - \mu_2(t)_{estimate}\|^2$

i.e. $T_n = \int n\,(\mu_1(t)_{estimate} - \mu_2(t)_{estimate})^2\,dt$

where, it is known that

$\mu_1(t)_{estimate} - \mu_2(t)_{estimate} \sim GP(\mu_1(t) - \mu_2(t), \frac{\gamma_1(s,t)}{n_1} + \frac{\gamma_1(s,t)}{n_2})$

T_n follows mixed chi-squared distribution whose degrees of freedom are calculated using calculus method and consistency.

Additional Tools

1. Bootstrapping techniques

Bootstrap techniques are relatively new to the field of statistics. In statistics, Bootstrapping means use of random sampling with replacement for estimation or testing in repetitive manner. The basic idea of bootstrapping is that, inference about a population from sample data can be modelled by repeatedly sampled data and performing inference about a sample from resampled data. In bootstrap-resamples, the 'population' is in fact the sample.

These techniques are developed to be used in both the above data analysis for zero inflation as well as functional data.

2. R-Software

R-software is an important statistical software. It is a free-ware and can be improved every now and then when necessary. This proves its usefulness in data analysis. Undoubtedly, it is used in model fitting and functional data analysis. From package **pscl**, zero inflated regression models are obtained using functions **hurdle( )**, **zeroinfl( )**. The function **glm( )** also helps in model fitting. **ZIM** package also contains various commands for the analysis. For functional data analysis, packages like **fda.usc**, **rainbow**, **ftsa**, **re-fund**, etc. are used for different purposes of the user.

The features of R-software change dynamically, new packages are often added for better results.

References

i. https://en.wikipedia.org

ii. https://stats.stackexchange.com

iii. https://cran.r-project.org

iv. *Score Tests for Zero-Inflated Poisson Models* by N. Jansakula from Department of Mathematics, Prince of Songkla University, Hatyai, Songkla 90112, Thailand and J.P. Hinde from School of Mathematical Sciences, Laver Building, Exeter University, North Park Road, Exeter EX4 4QE, UK.
Received 1 January 2001; received in revised form 1 November 2001; accepted 1 November 2001

v. *A Comparison of Two Test Statistics for Poisson Overdeposition / Underdeposition* by Hongyue Wang, Changyong Feng, Xinming Tu from Department of Biostatistics and Computational Biology, University of Rochester, Rochester, US and Jeanne Kowalski from Department of Biostatistics & Bioinformatics, Rollins School of Public Health Winship Cancer Institute, Emory University, Atlanta, USA.
Received May 1, 2012; revised June 1, 2012; accepted June 8, 2012

vi. *Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial* by Mei-Chen Hu, Ph.D.; Martina Pavlicova, Ph.D. and Edward V. Nunes, M.D.

vii. *Two Samples Tests for Functional Data* by Chongqi Zhang from Department of Probability and Statistics, Guangzhou University, Guangzhou, China; Heng Peng from Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, And Jin-Ting Zhang from Department of Statistics and Applied Probability, National University of Singapore, Singapore

viii. *Regression Models for Count Data in R* by Achim Zeileis from Universit¨at Innsbruck; Christian Kleiber Universit¨at Basel and Simon Jackman from Stanford University