# Modelling Small Sample Discrete Data Using Exact Distributions

**Rochita Das**

Data Science, SSP, PLS, Novartis Business Services

## Introduction

- Asymptotic methods are unreliable for analyzing **small, skewed, highly stratified or sparse data** sets; as the usual MLE may not be valid or the estimated dispersion matrix may be unbounded.
- Exact conditional inference, i.e. Inference based on enumerating the exact distributions of sufficient statistics for parameters of interest conditional on the remaining parameters, is valid in such situations.

## Objectives

- To show conclusions obtained from large sample approximation are seemingly conflicting and how exact analysis helps to overcome it.
- The advantages of using exact Logistic and Poisson distributions in different examples will be explored.

## Methodology

❑ The **exact logistic inference** using conditional maximum likelihood (**CML**) estimation method (Cyrus et. al., 1995) is described below:

- **Step 1:** Exact estimates of parameter(s) are derived by considering all other parameters as nuisance, which can be removed from the analysis by conditioning on their sufficient statistics to create the conditional likelihood.
  - sufficient statistics for the $\beta_j$ is $\mathbf{T_j} = \sum_{i=1}^{n} \mathbf{y_i}\, \mathbf{x_{ij}}$ , $T = (T_1, \ldots, T_p)'$
  - **Pr(T=t)** = $\frac{C(t)\, exp(t'\beta)}{\prod_{i=1}^{n}[\,1+exp(\,x_i'\beta)]}$ , C(t) is number of sequences of y that generate t.
  - Let, $T_0$ be sufficient statistics for the nuisance parameters $\beta_0$ with observed value as $t_0$ and corresponding column of X be $X_0$. Define $T_1$, $t_1$, and $X_1$ for the parameters of interest $\beta_1$.
  - Conditional likelihood, **Pr** $(T_1=t_1 \mid T_0=t_0) = \frac{C(t)\, \exp(\,t_1'\beta_1)}{\sum_u C(u,t_0)\, \exp(u'\beta_1)}$

where $C(u,t_0)$ is the number of vectors y such that $y'X_1 = u$ and $y'X_0 = t_0$

- **Step 2:** Estimation is performed by maximizing the conditional likelihood. **Newton-Raphson** algorithm is used to perform this search.
  - However, if conditional pdf is monotonically increasing in parameter then **Median Unbiased Estimate** (MUE) is used which satisfies $f_{\hat{\beta}_i}(t_i|t_0) = \frac{1}{2}$ .
  - To generate this conditional distribution for a large number of observations, **Multivariate Shift Algorithm** is used, which is described as follow:

✓ Let $y_{(i)} = (y_{1,\ldots,} y_i)'$, $X_{(i)} = (x_{1,\ldots,} x_i)'$

✓ Sufficient statistic based on these i rows $\mathbf{t}'_{(i)} = \mathbf{y}'_{(i)}\, \mathbf{X}_{(i)}$

✓ A recursion relation results: $\mathbf{t}_{(i+1)} = \mathbf{t}_{(i)} + \mathbf{y}_{i+1}\, \mathbf{x}_{(i+1)}$

**Table1: Example of Algorithm**

| Obs | Y | $x_0$ | $x_1$ |
|---|---|---|---|
| 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 |
| 3 | 0 | 1 | 2 |
| 4 | 1 | 1 | 0 |

**Fig 1: Representation of Algorithm**

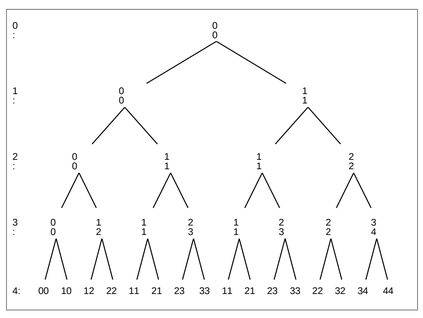


**Table 2: Distributions of sufficient statistics**

| Obs | $t_0$ | $t_1$ | Frequency | Probability |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1/16 |
| 2 | 1 | 0 | 1 | 1/16 |
| 3 | 1 | 1 | 2 | 2/16 |
| 4 | 1 | 2 | 1 | 1/16 |
| 5 | 2 | 1 | 2 | 2/16 |
| 6 | 2 | 2 | 2 | 2/16 |
| 7 | 2 | 3 | 2 | 2/16 |
| 8 | 3 | 2 | 1 | 1/16 |
| 9 | 3 | 3 | 2 | 2/16 |
| 10 | 3 | 4 | 1 | 1/16 |
| 11 | 4 | 4 | 1 | 1/16 |
| Total | | | 16 | 1 |

- **Step 3:** The joint and separate hypotheses are tested using **exact probability test** and **exact conditional score** test.
  - To test $\beta_1=0$ vs at least one element of $\beta_1$ is not 0
  - P-value, **p=** $\sum_{v \in E} f(v|\,\beta_1 = 0)$ ; E being critical region
  - Conditional probability test: $E_{cp}$ **={v: f(v** $|\,\beta_1=0) \le$ **f(**$t_1|\beta_1=0$**)}**
  - Conditional score test : $E_{cs} = \{\, v: (v-\mu_1)'\Sigma_1^{-1}(v-\mu_1) \ge (t_1-\mu_1)'\Sigma_1^{-1}(t_1-\mu_1)\}$

where $\mu_1$ is the mean and $\Sigma_1$ is the variance covariance matrix of $f(t_1|\beta_1=0)$.

❑ Similarly we proceed for **exact Poisson regression.**

## Results : Exact Logistic Regression

**Table 3: Data**

| Obs | $x_1$ (Treatment) | $x_2$ (Gender) | y | count |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 |
| 2 | 0 | 1 | 0 | 2 |
| 3 | 1 | 0 | 1 | 8 |
| 4 | 1 | 1 | 1 | 21 |

Data consists of–**number of patients** (**response**;0=not cured,1=cured), for several combinations of **treatment** (0=placebo,1=trial treatment) and **gender** (0=female,1=male) (**Independent variables**)

### ➢ Asymptotic Method Analysis

**Logistic Model:**

$\text{Logit}(\Pi_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

**Convergence Status:**

✓ Complete separation of data points

✓ **MLE does not exist**

### ➢ Exact Method Analysis

**Table 4: Sufficient Statistics**

| Parameter | value |
|---|---|
| Intercept | 2 |
| $x_1$ (Treatment) | 0 |
| $x_2$ (Gender) | 2 |

✓ Sufficient statistic lies at an extreme of the derived distribution, implying that **CMLE does not exist**

✓ Median unbiased estimate **(MUE) is used** instead of CMLE

**Table 5: Parameter Estimation**

| Parameter | Estimate | Standard error | 95% confidence limits | | p-value |
|---|---|---|---|---|---|
| $x_1$ (Treatment) | -3.84 * | - | - Infinity | -1.07 | 0.007 |
| $x_2$ (Gender) | 0.69 * | - | -2.97 | Infinity | 0.667 |

Note: * indicates a median unbiased estimate.

✓ Treatment ($x_1$) is significant

✓ Gender ($x_2$) is insignificant

✓ Joint exact test of $x_1$ and $x_2$ is significant

**Table 6: Testing of Hypothesis**

| Effect | Test | Statistic | P-value |
|---|---|---|---|
| Joint | Score | 21.115 | 0.002 |
| | Probability | 0.002 | 0.002 |
| $x_1$ (Treatment) | Score | 22 | 0.004 |
| | Probability | 0.003 | 0.004 |
| $x_2$ (Gender) | Score | 2 | 0.333 |
| | Probability | 0.333 | 0.333 |

## Results : Exact Poisson Regression

**Table 7: Data**

| | Soak-1 | Soak-1.7 | Soak-2.2 | Soak-2.8 | Soak-4 |
|---|---|---|---|---|---|
| Heat-7 | 0 (10) | 0 (17) | 0 (7) | 0 (12) | 0 (9) |
| Heat-14 | 0 (31) | 0 (43) | 2 (33) | 0 (31) | 0 (19) |
| Heat-27 | 1 (56) | 4 (44) | 0 (21) | 1 (22) | 1 (16) |
| Heat-51 | 3 (130 | 0 (1) | 0 (1) | 0 (1) | - |

Data consists of - the number of ingots that are not ready for rolling (**response**), out of Total tested (**offset**), for several combinations of heating time and soaking time (**categorical predictors**).

### ➢ Asymptotic Method Analysis

**Poisson Model:**

$\text{Log}(\lambda_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \text{Log}(t_i) + \varepsilon_i$

$t_i$=Total tested (offset variable)

**Table 8: Parameter Estimation**

| Parameter | Estimate | Standard error | 95% confidence limits | | p-value |
|---|---|---|---|---|---|
| Intercept | -1.57 | 1.16 | -3.85 | 0.71 | 0.18 |
| Heat-7 | -27.61 | 264324.6 | -518094 | 518039 | 0.99 |
| Heat-14 | -3.01 | 1 | -4.97 | -1.04 | 0.002 |
| Heat-27 | -1.72 | 0.77 | -3.22 | -0.21 | 0.02 |
| Soak-1 | -0.24 | 1.14 | -2.49 | 2 | 0.83 |
| Soak-1.7 | 0.56 | 1.12 | -1.64 | 2.75 | 0.62 |
| Soak-2.2 | 0.41 | 1.22 | -1.99 | 2.81 | 0.74 |
| Soak-2.8 | -0.13 | 1.42 | -2.92 | 2.66 | 0.93 |

✓ Standard errors reflects **convergence difficulties** for **Heat=7** parameter which implies that the parameter is **not estimable** by this method.

### ➢ Exact Method Analysis

✓ In the exact analysis, a **median unbiased estimate** is computed for the parameter **Heat=7** instead of a maximum likelihood estimate.

**Table 9: Parameter Estimation**

| Parameter | Estimate | Standard error | 95% confidence limits | | p-value |
|---|---|---|---|---|---|
| Heat-7 | -2.75 * | - | -Infinity | -0.79 | 0.02 |
| Heat-14 | -3.02 | 1.01 | -5.74 | -0.62 | 0.01 |
| Heat-27 | -1.78 | 0.81 | -3.68 | 0.23 | 0.08 |
| Soak-1 | -0.32 | 1.17 | -2.87 | 3.67 | 1 |
| Soak-1.7 | 0.53 | 1.13 | -1.81 | 4.46 | 1 |
| Soak-2.2 | 0.40 | 1.23 | -2.58 | 4.50 | 1 |
| Soak-2.8 | -0.17 | 1.42 | -4.55 | 4.22 | 1 |

Note: * indicates a median unbiased estimate.

**Table 10: Testing of Hypothesis**

| Effect | Test | Statistic | P-value |
|---|---|---|---|
| Joint | Score | 18.366 | 0.013 |
| | Probability | 1.29E-6 | 0.047 |
| $x_1$ (Heat) | Score | 15.825 | 0.002 |
| | Probability | 0.0002 | 0.006 |
| $x_2$ (Soak) | Score | 1.461 | 0.868 |
| | Probability | 0.007 | 0.817 |

✓ Heat and Soak are jointly significant

✓ Heat parameters conditional on Soak explain a significant amount of the variability

✓ Soak parameters conditional on Heat are not significant.

## Conclusion

➢ Thus it is recommended to use exact regression techniques in discrete small sample data or when the data is highly stratified.

➢ Further extensions can be done for multivariate regression and other univariate discrete distributions like negative binomial etc.

## References

1. Robert E. Derr, "Performing Exact Logistic Regression with the SAS System- Revised 2009"
2. Stokes, M. E., Davis, C. S., and Koch, G. G. (2000), "Categorical Data Analysis Using the SAS System"

NOVARTIS