Heterogeneous COVID-19 Incidences in Diamond Princess cruise ship: Some Volatile Informatics

Ramalingam Shanmugam
Honorary Professor of International Studies, School of Health Administration
Texas State University, San Marcos, TX 78666, USA.
ORCID 0000-0002-3388-1014, Email: rs25@txstate.edu

Abstract: This article centers around heterogeneity. It is popularly and briefly mentioned in literature but is not well defined, explained enough on how to compute or interpret. We here offer a formal definition and properties of heterogeneity. We propose an identifier of its existence. Other related ideas are discussed in this article. The concepts and new expressions of this article are illustrated using the COVID-19 data in a Diamond Princess cruise ship, during February 2020.

Key words: observed and non-observed homogeneity, Tango index, over/under dispersion, Poisson and binomial distributions.

1. Motivation.

What is heterogeneity, an antonym of homogeneity? Its root word lies in Greek “heterogenes” meaning different. In elementary through advanced statistics courses, the word heterogeneity is popularly commented to have existed when the
sample variance \( s^2 \) is large. In insurance applications, for example, the premium is assessed more if the insurer is heterogeneous compared to others and has hazard rate (Spreeuw, 1999). Should a large (small) sample variance be indicative of a homogeneity? See Ecochard (2006) for details. In healthcare disciplines, heterogeneity is referred in patient’s outcome level sometimes or in the latent genetic level in other times. Should the heterogeneity be connected to a non-observable hidden trait? Does the heterogeneity refer dissimilar attributes across the subgroups of the population itself even before sampling? Is heterogeneity really pointing out the non-identical nature in a random sample or population? Should heterogeneity imply a shifting population? In genetic studies, several authors refer that genetic heterogeneity is rather too difficult to ascertain. What does it really mean? If alleles in more than one locus exhibit susceptibility to a disease, there is a need to track the loci to infer their heterogeneity. So, in a sense, the application of heterogeneity is really a discussion of similarity or differences across loci (see Elston et al., 2003, pages 3404-344 for details). To realize how the heterogeneity plays a role to judge the crime victimization, see Hope and Norris (2013).

In the statistics literature, using a random sample \( y_1, y_2, \ldots, y_n \) from a population whose main parameter is \( \theta \), when the null hypothesis \( H_0: \theta_1 = \theta_2 = \ldots = \theta_n \) is tested, it is named \textit{homogeneity test}. This suggests that heterogeneity is really all about a shifting population. The students of statistics are more confused. Is the source of such confusion with respect to heterogeneity situated in its structure or in its ill-communication? It is evident that there is a lack of clear definition of heterogeneity (see chapter 12 of Hunink et al., 2018 for details). Neither the \textit{Encyclopedia of}
It is known that a binomial population possesses an under dispersion (that is, variance of the binomial distribution is smaller than its mean). In a Poisson population, the drawn random sample ought to reflect equality between the mean and variance. When the main (incidence rate) parameter of a Poisson chance mechanism is stochastically transient, the unconditional observation of the random variable convolutes to an inverse binomial model (Ross, 2002). The inverse binomial distribution is known to attest that the variance is larger than its mean. See Stuart and Ord (2015) for details. Consequently, a comparison between the mean and variance characterizes only which type binomial, Poisson, or inverse binomial is the underlying chance mechanism we are sampling from but not anything about the heterogeneity.

See Shanmugam (2020) for details about the probabilistic patterns among coronavirus confirmed, cured and deaths in the thirty-two India’s states/territories. To be concise about the confusion with respect to the heterogeneity, let us consider the data (of Mizumoto and Chowell, 2020) in Table 1 below describing the spread of COVID-19 among the voyagers in a Diamond Princess cruise ship, during February 2020. The random variables \( Y_1, Y_2, \) and \( Y_3 \) denote respectively the number of COVID-19 cases, the number of asymptomatic and symptomatic among them in a date. Under a given COVID-19’s prevalence rate, \( \lambda > 0 \), the number \( Y_1 \) perhaps follows a Poisson probability pattern. For a given number of COVID-19 cases in a date, the number \( Y_2 \) perhaps follows a binomial probability pattern with parameters
\((y_1, p)\), where \(0 < p < 1\) denotes the chance for a COVID-19 case to exhibit no symptom. Naturally, the number \(Y_i\) should follow a binomial probability pattern with parameters \((y_1, 1 - p)\). There is an implicitness between \(Y_2\) and \(Y_3\) in the sense that \(Y_2 + Y_3 = Y_i\). There are three-time oriented groups of COVID-19 incidences in Table 1. Is there an observable heterogeneity among the three groups? If so, is it due to a non-observable (parametric) heterogeneity? How do we define and distinguish observable versus non-observable heterogeneity? A literature search to answer this question is in vain.

Table 1. COVID-19 in cruise ship, 2020, Mizumoto et al. (2020)

| Date       | \(Y_3\) | \(Y_2\) | \(Y_1\) | \(\bar{\lambda} = \bar{y}_1\) | \(s^2_{\bar{\lambda}} = \text{Var}(\bar{\lambda})\) | \(OR_{\bar{\lambda}}\) | \(Odds_{\bar{\lambda}}\) |
|------------|---------|---------|---------|-------------------------------|-------------------------------|----------------|-----------------|
| 15, 16 Feb | 29, 32  | 38, 38  | 67, 70  | 68.5                          | 4.5                           | 0.5001         | 1.7E-30         |
| 2020       |         |         |         |                               |                               |                |                 |
| 17, 18 Feb | 29, 23  | 70, 65  | 99, 88  | 93.5                          | 60.5                          | 0.5000         | 2.4E-41         |
| 2020       |         |         |         |                               |                               |                |                 |
| 19, 20 Feb | 11, 7   | 68, 6   | 79, 13  | 46                            | 21.78                         | 0.5002         | 1.0E-20         |
| 2020       |         |         |         |                               |                               |                |                 |

It is evident that the average of the COVID-19 cases is estimate of the COVID-19’s prevalence rate (that is, \(\bar{\lambda}\) in Table 1). Their estimates impress that prevalence rate is transient, not constant across every pair of two days duration. The Poisson population from which the COVID-19 cases are drawn ought to have been dynamic, implying the existence of a Poisson heterogeneity.
Likewise, given that a fixed number, $y_1$ of COVID-19 cases has occurred, a part of them is asymptomatic cases, $y_2$ and the remaining are symptomatic cases, $y_3$. That is, $y_2$ and $y_3$ are complementary such that $y_2 + y_3 = y_1$. Is there heterogeneity in each of the two sub-binomial populations? How should each binomial heterogeneity be defined and computed? A literature search offers help to prove the existence or absence of binomial heterogeneity in the data for $y_2$ or $y_3$ in Table 1.

The concept of heterogeneity seems to have escaped the researchers and statisticians’ scrutiny for a long time. It is worthwhile to revive an interest in the heterogeneity and that is exactly what this article is trying to accomplish. Hence, we first define and construct an approach for the idea of heterogeneity. To be specific, we first discuss the Poisson heterogeneity and then take up the binomial heterogeneity. May be our research direction about the heterogeneity in this article is perhaps a pioneering type. However, we believe that our approach in this article is easily extendable for many other similar distributional set up. In the end, we illustrate our definition and all derived expressions for the heterogeneity in this article using the COVID-19’s data pertaining to the Diamond Princess cruise ship, Yokohama, 2020 as displayed in the Table 1.

2. How to define Poisson and binomial heterogeneities?

Applied and theoretical statisticians emphasize that the heterogeneity is a paramount importance in extracting and interpreting data evidence. Many data analysts are convinced that an unrecognized heterogeneity leads to a biased inference. To begin with, what is heterogeneity? It is a factor causing non-
similarities. If so, are there two sources for heterogeneity to arise or exist? One source ought to be the drawn random sample of observations: \( y_1, y_2, \ldots, y_n \), which we recognize as observable heterogeneity. Would the sampling variability, 
\[ \text{Var}[f(y_1, y_2, \ldots, y_n | \theta)] \]  
for a selected statistic \( f(y_1, y_2, \ldots, y_n | \theta) \) express the observable heterogeneity? Another source is manifested in non-observable parameter, \( \theta \) of the chance mechanism, which we recognize as non-observable heterogeneity. Would a non-uniform stochastic pattern of \( \theta \) be indicative of the non-observable homogeneity? If the chance mechanism perversely selects a probability density function (pdf) for \( \theta \), how would it manifest itself to portray the non-observable heterogeneity? Both observable and non-observable heterogeneities together ought to be involved to make any definition of heterogeneity to be complete. If so, how do we integrate them? Often, under/over-dispersion is confused to be heterogeneity. It seems that the over/under dispersion is precipitated by the heterogeneity. It is not obvious or proven on whether the converse is true. In this article, we focus only on Poisson and binomial populations to address the heterogeneity and the arguments can be repeated for other populations.

### 2.1 Poisson heterogeneity.

Recall that the random integer, \( Y_i \), denoting the number of COVID-19 cases in a place (like the Diamond Princess cruise ship) at a time (like February, 2020) is a Poisson random variable with a specified prevalence rate, \( \lambda > 0 \). That is, the conditional probability of observing \( y_i \) number of COVID-19 cases under a prevalence rate \( \lambda > 0 \) is 
\[ \text{Pr}[Y_i = y_i | \lambda] = e^{-\lambda \lambda^y_i} / y_i !; y_i = 0, 1, 2, \ldots; \lambda > 0 \]
with its expected number \(E[Y_i|\lambda] = \lambda\) and variability \(Var(Y_i|\lambda) = E[Y_i|\lambda]\) (see the derivation details Poisson mean and variance in Rajan and Shanmugam, 2020). The prevalence parameter \(\lambda\) itself is crucial in our discussions. The Poisson variability cannot be the heterogeneity because the expected value also changes when the variability changes, due to their inter-relatedness. Realize that no two individuals in the ship are assumed to have same level of susceptibility to the COVID-19 virus. It is reasonable to imagine that the prevalence levels follow a conjugate, stochastic gamma distribution. The so-called conjugate prior knowledge in the Bayesian framework smooths the statistical process. It is known that the conjugate prior for the Poisson distribution is gamma, whose pdf is

\[
c(\lambda|\alpha, \beta) d\lambda = e^{-\alpha\lambda} (\alpha\lambda)^{\beta-1} / \Gamma(\beta); \alpha > 0; \beta > 0,
\]

with an average.

\[
E(\lambda|\alpha, \beta) = \frac{\beta}{\alpha}
\]

and variability \(Var(\lambda|\alpha, \beta) = E(\lambda|\alpha, \beta) / \alpha\), where the parameters \(\alpha\) and \(\beta\) are recognized as hyper-parameters (for details, see Rajan and Shanmugam, 2020). Notice that the hyper parameter \(\alpha > 0\) causes the variability in the COVID-19’s prevalence rate to fluctuate up or down and hence, you would anticipate the heterogeneity to involve the hyperparameter \(\alpha\). But the question is how?

Assume that the probability of observing a non-negative COVID-19 cases, \(y_i\) is a Poisson under a stable sampling population \(Pr(Y_i|\lambda)\) with an expected number \(E(Y_i|\lambda) = \lambda\) and a variability \(Var(Y_i|\lambda) = E(Y_i|\lambda)\). With replications, the observable heterogeneity should be estimable. That is, the maximum likelihood estimate (MLE) of the COVID-19 prevalence rate is the average number, \(\bar{y}_i\) of the
observations. To discuss the non-observable heterogeneity, we need to integrate its conjugate prior $c(\lambda | \alpha, \beta)$ for the non-observable $\lambda$ with the likelihood $\Pr(Y_i | \lambda)$ and it results in an update called posterior pdf for $\lambda$. Then, 
\[ c(\lambda | y_1, \alpha, \beta) = \Pr(y_1 | \lambda)c(\lambda | \alpha, \beta) / \int_{-\infty}^{\infty} \Pr(y_1 | \lambda)c(\lambda | \alpha, \beta)d\lambda \]
is the posterior pdf of the non-observable $\lambda$. Also, the denominator
\[ \int_{0}^{\infty} \Pr(y_1 | \lambda)c(\lambda | \alpha, \beta)d\lambda = \Gamma(\alpha + y_1) / (1 + \beta)^{\alpha + y_1} \]
in a Bayesian framework is called the marginal distribution. With $\Delta_\lambda = \lambda - E(\lambda)$, it is clear that $\int_{-\infty}^{\infty} \Delta_\lambda c(\lambda | \alpha, \beta)d\lambda = 0$
, note that the prior variance is $\text{Var}(\lambda | \alpha, \beta) = \int_{-\infty}^{\infty} \Delta_\lambda^2 c(\lambda | \alpha, \beta)d\lambda$. Because the prior is conjugate, its counterpart’s variability
\[ \text{Var}(\lambda | y_1, \alpha, \beta) = \int_{-\infty}^{\infty} (\lambda - E[\lambda | y_1, \alpha, \beta])^2 c(\lambda | \alpha, \beta)d\lambda \]
is minimal when the Bayes estimate of the non-observable is the posterior mean, $\lambda_{\text{Bayes}} = E[\lambda | y_1, \alpha, \beta]$, where $E[\lambda | y_1, \alpha, \beta] = (\alpha + \bar{y}) / \beta$. Differentiating the log-likelihood function
\[ \ln L(n\bar{y}, \lambda) = n\bar{y}_1 \ln \lambda - n\lambda + \sum_{i=1}^{n} \ln(y_i !) \]
with respect to the non-observable parameter, $\lambda$, setting it equal to zero and solving it, we obtain the MLE and it is $\hat{\lambda}_{\text{MLE}} = \bar{y}_1$. Because of the invariance property of the MLE, it is involved. The invariance property refers to that the MLE of a function of the parameter is the function of the MLE of the parameter. Also, it is known (Blumenfeld, 2010) that
\[ E_{\text{prior}} E_{\text{likelihood}}(\bar{y}_1 | \lambda) = E(\bar{y}_1) \]

and

\[ \text{Var}(\bar{y}_1) = E_{\text{prior}} \text{Var}_{\text{likelihood}}(\bar{y}_1 | \lambda) + \text{Var}_{\text{prior}} E_{\text{likelihood}}(\bar{y}_1 | \lambda). \]

Hence, we are ready now to define the non-observable heterogeneity below in the Definition 1.

**Definition 1.** The non-observable heterogeneity of the Poisson parameter, \( \lambda \) is defined as \( H_\lambda = [1 + \frac{\text{Var}_{\text{prior}} E_{\text{likelihood}}(\bar{y}_1 | \lambda)}{E_{\text{prior}} \text{Var}_{\text{likelihood}}(\bar{y}_1 | \lambda)}]^{-1} \in [0,1]. \)

Following the Definition 1, we obtain the *non-observable heterogeneity* of the COVID-19 cases is \( H_\lambda = [1 + \frac{1}{\beta}]^{-1} \in [0,1]. \) When the value of \( H_\lambda \) is closer to zero, the data are believed to have non-observable Poisson homogeneity. Its MLE is

\[ \hat{H}_\lambda = [1 + \frac{1}{\hat{\beta}}]^{-1} = [1 + \frac{s^2_\lambda}{\lambda}]^{-1} = \frac{\lambda}{\hat{\lambda} + s^2_\lambda}. \]

See Figure 1 for the configuration of the non-observable Poisson heterogeneity in general. Likewise, the *observable heterogeneity* is defined below in Definition 2.

**Definition 2.** The observable heterogeneity of the randomly sampled Poisson counts, \( y_1, y_2, \ldots, y_n \) is defined as \( H_{y_1} = [1 + \frac{\text{Var}_{\text{marginal}} E_{\text{posterior}}(\lambda | \bar{y}_1)}{E_{\text{marginal}} \text{Var}_{\text{posterior}}(\lambda | \bar{y}_1)}]^{-1} \in [0,1]. \)

Before we apply the Definition 2, let us recollect that the marginal pdf of the complete sufficient statistic, \( \bar{y}_1 \) is uniform distribution and the posterior
distribution is 
\[ c(\lambda \mid \bar{y}_i, \alpha, \beta) = (1 + \beta)^{(\alpha + n\bar{y}_i)} \left[ e^{-\left(1 + \beta\right)\lambda} \right]^{(\alpha + n\bar{y}_i)-1} / \Gamma(\alpha + n\bar{y}_i) \] 
with

\[ E(\lambda \mid \bar{y}_i, \alpha, \beta) = \frac{(\alpha + n\bar{y}_i)}{(1 + \beta)} \]
and
\[ \text{Var}(\lambda \mid \bar{y}_i, \alpha, \beta) = \frac{E(\lambda \mid \bar{y}_i, \alpha, \beta)}{(1 + \beta)}. \]

Imposing the Definition 2 and simplifying, we obtain that \( H_{\bar{y}_i} = [1 + \frac{(1 + \beta)}{6}]^{-1} \) whose MLE is

\[ \hat{H}_{\bar{y}_i} = [1 + \frac{(1 + \hat{\beta})}{6}]^{-1} = [1 + \frac{(1 + \frac{\hat{\lambda}}{s^2_\lambda})}{6}]^{-1} \in [0,1]. \]

See Figure 2 for the configuration of the observable Poisson heterogeneity, \( \hat{H}_{\bar{y}_i} \) in general. When the value of \( \hat{H}_{\bar{y}_i} \) is closer to zero, the data are interpreted to have observable homogeneity.

\[ \text{Figure 1. Non-observable heterogeneity} \quad \text{Figure 2. Observable heterogeneity} \]

Furthermore, the distance, \( d(y_i, \hat{\lambda}) \) between the observable \( y_i \) of the number of COVID-19 cases and the prevalence rate \( \lambda \) could be assessed using the formula

\[ d(y_i, \lambda) = E_y E_{\lambda} \left| Y_i - \lambda \right| = \sum_{y_i=0}^{\infty} \int \left| Y_i - \lambda \right| \text{Pr}(y_i \mid \lambda) c(\lambda \mid \bar{y}_i, \alpha, \beta) d\lambda. \]

Realizing that their absolute difference is really \( \left| Y_i - \lambda \right| = Y_i + \lambda - 2 \min\{Y_i, \lambda\} \), we obtain after simplifications that \( d(y_i, \lambda) = \beta(1 - \beta) \pm 1 \left( \frac{\alpha}{1 + \beta} \right). \)

See Figure 3 for the
configuration of the distance, \(d(y_1, \lambda)\) between the observable and non-observable in Poisson mechanism.

We now turn to discuss stochastic properties of the Poisson distribution. The survival function of the random number, \(Y_1\) of COVID-19 cases is
\[
S_{Y_1}(r | \lambda) = \Pr(Y_1 \geq r | \lambda) = \sum_{i=r}^{\infty} e^{-\lambda} \lambda^i / i! = P[\chi^2_{2(r+1)df} < 2\lambda]; \lambda > 0.
\]
The hazard rate is a force of mortality. The hazard rate, \(h(y_1)\) for the COVID-19 occurrence is
\[
h(y_1) = \frac{P(Y_1 \geq r | \lambda)}{S(Y_1 + 1 | \lambda)} = \frac{e^{-\lambda} \lambda^y}{y! P[\chi^2_{2(y+1)df} < 2\lambda]; \lambda > 0}.
\] Does the Poisson chance mechanism keep any a finite memory? For example, the geometric distribution is known to have no memory. The memory is really a conditional probability. That is,
\[
\text{memory} = \Pr(Y_1 \geq s | Y_1 \geq r) = \frac{\Pr(Y_1 \geq r + s)}{\Pr(Y_1 \geq r)} = \frac{P[\chi^2_{2(r+s+1)df} < 2\lambda]}{P[\chi^2_{2(r+1)df} < 2\lambda]; \lambda > 0},
\]
confirming that there is a finite memory in the Poisson mechanism of COVID-19 incidences. To be specific, with \(r = 0, s = 1\) in the above result, the memory between COVID-19 free situation and just one COVID-19 occurrence is revealed in the chance oriented Poisson mechanism. Such a memory is
\[
\text{memory}_{0\rightarrow 1} = \frac{P[\chi^2_{4df} < 2\lambda]}{P[\chi^2_{2df} < 2\lambda]; \lambda > 0}.
\] Likewise, the memory between at least one COVID-19 case situation and at least two COVID-19 cases situation is revealed with a substitution of \(r = 1, s = 1\) in the above result and it is
\[
\text{memory}_{1\rightarrow 2} = \frac{P[\chi^2_{6df} < 2\lambda]}{P[\chi^2_{4df} < 2\lambda]; \lambda > 0}.
\] The odds ratio from the initial memory to the next memory is
\[
OR_{1\rightarrow 2}^{0\rightarrow 1} = \frac{P[\chi^2_{6df} < 2\lambda]P[\chi^2_{2df} < 2\lambda]}{P[\chi^2_{4df} < 2\lambda]} (\text{see})
their values in Table 1). However, the odds for COVID-19 free healthy situation to prevail is $Odds_{Y_i} = \frac{\Pr(Y_i = 0)}{\Pr(Y_i \geq 1)} = (e^\lambda - 1)^{-1}; \lambda > 0$ (see their values in Table 1). For details on how the chance for an incidence of a disease to occur from a disease-free scenario changes, see Shanmugam and Radhakrishnan (2011).

2.2 Binomial heterogeneity

In this section, we explore the heterogeneity for two sub binomial processes emanating from a Poisson process. The asymptomatic number, $Y_2$ and symptomatic number, $Y_3$ of COVID-19 cases are two branching binomial random numbers out of the Poisson random number, $Y_1 = 0,1,2,...; of COVID-19 cases. These two split random variables are complementary of each other in the sense that $Y_2 + Y_3 = Y_1$. Then, what are the underlying model for $Y_2$ and for $Y_3$? Are they correlated random variables? If so, what is their correlation? These are pursued in this section.

Let an indicator random variable, $I_i$=1 for a COVID-19 person to show asymptomatic symptom with a probability, $0 < p < 1$ and $I_i$ = 0 to show symptomatic symptom with a probability, $0 < 1 - p < 1$. Then, for a fixed $y_1$, the random variable, $Y_2 = \sum_{i=1}^{y_1} I_i$ follows a binomial probability distribution with parameters $(y_1, p)$. Likewise, for a fixed $y_1$, the random variable, $Y_3 = y_1 - Y_2$ follows a complementary binomial distribution with parameters $(y_1, 1 - p)$. That is, $\Pr(Y_2 = y_2 \mid y_1, p) = \binom{y_1}{y_2} p^{y_2} (1 - p)^{y_1 - y_2}; y_2 = 0,1,2,......, y_1; 0 < p < 1$
\[
\Pr(Y_3 = y_3 \mid y_1, p) = \left( \frac{y_1}{y_3} \right)^\frac{1}{\rho} \left( 1 - \frac{y_1}{y_3} \right)^{y_3 - y_1}; y_3 = 0, 1, 2, \ldots, y_1; 0 < 1 - p < 1
\]

with their conditional expected numbers \( E(Y_2 \mid y_1, p) = y_1 p \), \( E(Y_3 \mid y_1, 1 - p) = y_1 (1 - p) = y_1 - E(Y_2 \mid y_1, p) \) and the conditional variabilities \( \text{Var}(Y_2 \mid y_1, p) = (1 - p) E(Y_2 \mid y_1, p) \) and \( \text{Var}(Y_3 \mid y_1, 1 - p) = p E(Y_3 \mid y_1, 1 - p) \). The conditional variability of \( Y_2 \) is a percent \((1 - p)\) of its expected number \( E(Y_2 \mid y_1, p) \), implying that it exhibits under dispersion. Likewise, the conditional variability of \( Y_3 \) is a percent \((1 - p)\) of its expected number \( E(Y_3 \mid y_1, p) = y_1 (1 - p) \), implying that it also exhibits under dispersion. Together, the above statements suggest a conditional balance \( \frac{E(Y_2 \mid y_1, p)}{E(Y_3 \mid y_1, 1 - p)} = \text{odds(asymptomatic)} = \frac{p}{(1 - p)} \) (See Stuart and Ord, 2015 for details of the odds concepts). Consequently, we note that

\[
p = \frac{E(Y_2 \mid y_1, p)}{E(Y_2 \mid y_1, p) + E(Y_3 \mid y_1, 1 - p)}.
\]

Furthermore, we wonder whether the random variables \( Y_2 \) and \( Y_3 \) are correlated? The answer is affirmative. To identify their correlation, notice that

\[
E(Y_2) = E_{y_1} E(Y_2 \mid y_1) = E_{y_1} (Y_2 p) = p \lambda,
\]

\[
E(Y_3) = E_{y_1} E(Y_3 \mid y_1) = E_{y_1} (Y_1 (1 - p)) = (1 - p) \lambda
\]

\[
\text{Var}(Y_2) = E_{y_1} \text{Var}(Y_2 \mid y_1) + \text{Var}_{y_1} E(Y_2 \mid y_1) = E_{y_1} \{Y_2 p (1 - p)\} + \text{Var}_{y_1} (Y_2 p) = p \lambda
\]

\[
\text{Var}(Y_3) = E_{y_1} \text{Var}(Y_3 \mid y_1) + \text{Var}_{y_1} E(Y_3 \mid y_1) = E_{y_1} \{Y_1 p (1 - p)\} + \text{Var}_{y_1} (Y_1 (1 - p)) = (1 - p) \lambda
\]
\[ \text{Cov}(Y_2, Y_3) = E_{y_1}E(Y_2|Y_3 \mid y_1) - E_{y_1}E(Y_2 \mid y_1)E_{y_1}E(Y_3 \mid y_1) \]

where

\[ E_{y_1}E(Y_2 \mid y_1) = p\lambda \text{, } E_{y_1}E(Y_3 \mid y_1) = (1-p)\lambda \text{,} \]

\[ E_{y_1}E(Y_2Y_3 \mid y_1) = E_{y_1}E_{y_2}E_{y_3|y_1} \{ Y_2 Y_3 \mid Y_1 \} = E_{y_1}E_{y_2}Y_2 \{ Y_2 (1-p) \} = E_{y_1} \{ p(1-p)Y_1^2 \} = p(1-p)\lambda(1+\lambda) \]

Hence, their correlation is \[ \rho_{y_2,y_3} = \frac{\text{Cov}(Y_2, Y_3)}{\sqrt{\text{Var}(Y_2)\text{Var}(Y_3)}} = \sqrt{p(1-p)} \text{.} \] Their expected distance, \[ d(Y_2, Y_3) = E_{y_1}E(\lvert Y_2 - Y_3 \rvert \mid Y_1) \text{ portray the drift between the symptomatic observable, } Y_2 \text{ and the asymptomatic observable, } Y_3 \text{ and it is simplified to } \]

\[ d(Y_2, Y_3) = \lvert 2p - 1 \rvert \lambda \text{ (see Table 3 for their values), due to applying } \]

\[ \lvert Y_2 - Y_3 \rvert = Y_2 + Y_3 - 2 \min\{Y_2, Y_3\} \text{.} \]

Let us assume that every COVID-19 case has the same chance of being asymptomatic in a period of time. Then, the random number, \( y_2 \) for a specified number, \( y_1 \) of COVID-19 cases follows a binomial distribution with parameters \( (y_1, p) \). We select a conjugate beta prior distribution

\[ c(p|\gamma, \delta)dp = \Gamma(\gamma + \delta) p^{\gamma-1}(1-p)^{\delta-1} / \Gamma(\gamma)\Gamma(\delta); 0 < p < 1; \gamma, \delta > 0 \text{ for our discussion for asymptomatic COVID-19 cases. The prior average is } \]

\[ \mu_{\text{prior}} = E(p|\gamma, \delta) = \frac{\gamma}{(\gamma + \delta)} \text{ and the prior variability is } \]

\[ \text{Var}(p|\gamma, \delta) = \mu_{\text{prior}} (1 - \mu_{\text{prior}}) / (1 + \gamma + \delta) \text{, where the parameters } \gamma \text{ and } \delta \text{ are hyper-parameters (see Rajan and Shanmugam, 2020 for details). We guess that the} \]
binomial heterogeneity would involve both hyperparameters. The task for us is how
do we construct such heterogeneity? An answer is the following. The posterior
distribution

\[ c(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \Pr(\bar{y}_2|\bar{y}_1, p)c(p|\gamma, \delta)/\int_{-\infty}^{\infty} \Pr(\bar{y}_2|\bar{y}_1, p)c(p|\gamma, \delta)dp \]

\[ = p^{\delta \bar{y}_1 - \gamma - 1}(1-p)^{\delta \bar{y}_2 - 1}/\{\Gamma(\gamma + \bar{y}_2)\Gamma(\delta + \bar{y}_1 - \bar{y}_2)/\Gamma(\gamma + \delta + \bar{y}_1)\} \]

would play a key role to construct both the observable and non-observable binomial
heterogeneity. With \( \Delta_p = p - E(p) \), it is clear that \( \int_{-\infty}^{\infty} \Delta_p c(p|\gamma, \delta)dp = 0 \). The prior
variance is \( \text{Var}(p|\gamma, \delta) = \int_{-\infty}^{\infty} \Delta_p^2 c(p|\gamma, \delta)dp \). Its posterior counterpart

\[ \text{Var}(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \int_{-\infty}^{\infty} (p - E[p|\bar{y}_1, \bar{y}_2, \gamma, \delta])^2 c(p|\gamma, \delta)dp \]

is minimal when the
Bayes estimate of non-observable is the posterior mean \( p_{\text{Bayes}} = E(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) \),
where \( \mu_{\text{posterior}} = E(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \frac{(\gamma + \bar{y}_2)}{(\gamma + \delta + \bar{y}_1)} \). The posterior variance is

\[ \text{Var}(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \frac{\mu_{\text{posterior}} (1 - \mu_{\text{posterior}})}{(1 + \gamma + \delta + \bar{y}_1)} \]. Differentiating the log-likelihood
function \( \ln L(n, \bar{y}_1, \bar{y}_2, p) = \bar{y}_2 \ln p + (\bar{y}_1 - \bar{y}_2) \ln(1-p) + \sum_{i=1}^{n} \ln(y_{i}) \) with respect to
the non-observable parameter, \( p \), setting it equal to zero and solving it, we obtain
the MLE and it is \( \hat{p}_{\text{MLE}} = \frac{\bar{y}_2}{\bar{y}_1} \). It is known that \( E_{\text{prior}}E_{\text{likelihood}}(\bar{y}_2|\bar{y}_1, p) = E(\bar{y}_2) \) and

\[ \text{Var} (\bar{y}_2) = E_{\text{prior}}\text{Var}_{\text{likelihood}}(\bar{y}_2|\bar{y}_1, p) + \text{Var}_{\text{prior}}E_{\text{likelihood}}(\bar{y}_2|\bar{y}_1, p) \]. Hence, we define
the non-observable binomial heterogeneity below in Definition 3.
Definition 3. The non-observable binomial heterogeneity is defined as

\[
H_{\gamma_1,\delta} = \left[1 + \frac{Var_{\text{prior}} E_{\text{likelihood}}(\overline{y}_2 | \overline{y}_1, p)}{E_{\text{prior}} Var_{\text{likelihood}}(\overline{y}_2 | \overline{y}_1, p)}\right]^{-1} \in [0, 1].
\]

Following the Definition 3, we obtain the non-observable heterogeneity of the COVID-19’s asymptotic cases (remembering that \((\overline{y}_1, \gamma, \delta)\) are the non-observable parameters) as

\[
H_{\gamma_1,\gamma,\delta} = \left[1 + y_1 \frac{Var_{\text{prior}}(p)}{E_{\text{prior}}(p(1 - p))}\right]^{-1} = \left[1 + \frac{y_1}{(\gamma + \delta)(1 + \gamma + \delta)}\right]^{-1} \in [0, 1].
\]

When the value of \(H_{\gamma_1,\gamma,\delta}\) is closer to zero, the data are interpreted to have non-observable binomial homogeneity. Substituting the MLEs \(\hat{\gamma} = \bar{p}\left(\frac{\bar{p}(1 - \bar{p})}{s^2_p} - 1\right)\) and \(\hat{\delta} = \left(1 - \bar{p}\right)\hat{\gamma}\), we obtain its MLE \(\hat{H}_{\gamma_1,\gamma,\delta} = \left[1 + \frac{y_1(s^2_p)^2}{\left(\bar{p}(1 - \bar{p}) - s^2_p\right)\bar{p}(1 - \bar{p})}\right]^{-1} \in [0, 1].\)

Likewise, the observable heterogeneity of the binomial distribution of \(y_2\) is defined below in the Definition 4.

Definition 4. The observable heterogeneity of the binomial counts, \(y_{2,i}, i = 1, 2, ..., y_1\) (in terms of the complete sufficient statistic \(\overline{y}_2\)) is defined as

\[
H_{y_2} = \left[1 + \frac{Var_{\text{marginal}} E_{\text{posterior}}(p | \overline{y}_1)}{E_{\text{marginal}} Var_{\text{posterior}}(p | \overline{y}_1)}\right]^{-1} \in [0, 1].
\]
Before we apply the Definition 4, remember that the marginal pdf of the complete sufficient statistic, \( \bar{y}_2 \) is beta-binomial distribution, 

\[
\Pr(\bar{y}_2) = \binom{\bar{y}_1}{\bar{y}_2} \frac{\Gamma(\gamma + \bar{y}_2)\Gamma(\delta + \bar{y}_1 - \bar{y}_2)}{\Gamma(\gamma + \delta + \bar{y}_1)}; \text{ and the posterior distribution is beta. With the notation } B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}, \text{ we note that the probability mass function of the beta-binomial distribution is}
\]

\[
\Pr(y_2) = \binom{y_1}{y_2} B(y_2, \gamma, \delta + y_1 - y_2) / B(\gamma, \delta); y_2 = 0, 1, \ldots, y_1; \gamma, \delta > 0.
\]

That is, the posterior probability density function is 

\[
c(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \frac{\Gamma(\gamma + \delta + \bar{y}_1)}{\Gamma(\gamma + \bar{y}_2)\Gamma(\delta + \bar{y}_1 - \bar{y}_2)} p^{\gamma + \bar{y}_2 - 1}(1 - p)^{\delta + \bar{y}_1 - \bar{y}_2 - 1}
\]

with 

\[
E(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \frac{(\gamma + \bar{y}_2)}{([\gamma + \delta] + \bar{y}_1)}
\]

and 

\[
Var(p|\bar{y}_1, \bar{y}_2, \gamma, \delta) = \frac{(\gamma + \bar{y}_2)(\delta + \bar{y}_1 - \bar{y}_2)}{([\gamma + \delta] + \bar{y}_1)(1 + [\gamma + \delta] + \bar{y}_1)}.
\]

Now applying the Definition 4, we obtain an expression for the observable binomial heterogeneity

\[
H_{y_2|y_1} = [1 + \frac{Var\text{marginal}\left\{ \frac{\gamma + y_2}{\gamma + \delta + y_1} \right\}}{E\text{marginal}\left\{ \frac{\gamma + y_2}{\gamma + \delta + y_1} \right\}}]^{-1} \approx 1 + \frac{\delta}{\gamma + \delta}(-\frac{y_1}{y_1 + \delta})^{-1} \in [0, 1],
\]

whose estimate is 

\[
\hat{H}_{y_2|y_1} \approx [1 + \hat{\mu}(\frac{y_1 s_p^2}{y_1 s_p^2 + (1 - \hat{\mu})(\hat{\mu}(1 - \hat{\mu}) - s_p^2)})]^{-1}, \text{because}
\]

\[
\hat{\gamma} \approx \hat{\mu}(\frac{\bar{y}(1 - \bar{y})}{s_p^2} - 1) \text{ and } \hat{\delta} \approx \hat{\gamma}(1 - \bar{y}) / \bar{y}.
\]

When the value of \( \hat{H}_{y_2|y_1} \) is closer to zero, the data are considered to have observable binomial homogeneity. Also, the distance,
\( d(y_2, p) \) between the observable \( y_2 \) of the number of asymptomatic COVID-19 cases and its proportion, \( p \) could be assessed using the formula

\[
d(y_2, p) = E_{y_2, p} |Y_2 - p| = \sum_{y_2 = 0}^{\infty} \int \left| Y_2 - p \right| \Pr(y_2 | p) c(p | y_1, y_2, \gamma, \delta) dp .
\]

Realizing that the absolute difference, \( |Y_2 - p| = Y_2 + p - \min\{Y_2, p\} \), we obtain after simplifications that \( d(y_2, p) = |y_1 - 1| \left( \frac{\gamma}{\gamma + \delta} \right) \).

Likewise, to obtain the non-observable heterogeneity of the COVID-19’s symptomatic cases, all we have to do is change \( p \) to \((1 - p)\), change \( y_2 \) to \( y_3 \), along with changing \( \gamma \) to \( \delta \) and go through the process above. Hence, the non-observable heterogeneity in the symptomatic cases is the same. That is,

\[
H_{y_3|y_3} \approx [1 + \left( \frac{\gamma}{\delta + \gamma} \right) \left( \frac{y_1}{y_1 + \gamma} \right)]^{-1} \in [0,1],
\]

whose MLE is

\[
\hat{H}_{y_3|y_3} \approx [1 + \bar{p} \left( \frac{\gamma y_3}{y_1 \gamma^2 + \bar{p} (1 - \bar{p}) - \gamma^2} \right)]^{-1},
\]

which is interestingly not the same as

\[
\hat{H}_{y_2|y_1}. \]

Also, the distance, \( d(y_3, 1 - p) \) between the observable \( y_3 \) of the number of asymptomatic COVID-19’s symptomatic cases and the proportion, \( 1 - p \) could be assessed using the formula
\[ d(y, 1 - p) = E_y E_p |Y - (1 - p)| = \sum_{y_{\text{as}} = 0}^{\infty} \left| Y - (1 - p) \right| \Pr(y, 1 - p) c(1 - p | y_1, y_2, \gamma, \delta) d(1 - p) \]

and it is after simplifications that \( d(y, p) = |y - 1| \left( \frac{\delta}{\gamma + \delta} \right) \).

Now we explore statistical properties of the asymptomatic cases, \( y_2 \). The survival function of the random number, \( Y_2 \) with asymptotic symptoms is

\[ S_{y_2}(r, p | y_1) = \Pr(Y_2 \geq r | y_1) = \sum_{i=r}^{\infty} \frac{y_1!}{i!(y_1 - i)!} p^i (1 - p)^{y_1 - i} \]

\[ = P \left[ F_{(2r, 2[y_1 - r + 1])df} \leq \frac{y_1 p(y_1 - r + 1)}{(1 - p)r} \right]; 0 < p < 1. \]

The hazard rate, \( h(y) \) of the binomial distribution for the asymptomatic cases is

\[ h(y_2) = \frac{\Pr(y_2 | p)}{S(y_2 + 1 | p)} = \frac{y_1! \left( \frac{p}{1 - p} \right)^{y_2} (1 - p)^{y_1}}{y_2! (y_1 - y_2)! \left[ F_{(2[r, 2[y_1 - r + 1])df} \right]} \leq \frac{y_1 p(y_1 - r + 1)}{(1 - p)r} \]; 0 < p < 1. \]

The binomial distribution has a finite memory

\[ \Pr(Y_2 \geq s | y_2 \geq r) = \frac{\Pr(Y_2 \geq r + s)}{\Pr(Y_2 \geq r)} = \frac{P \left[ F_{(2r, 2[y_1 - (r + s) + 1])df} \leq \frac{y_1 p(y_1 - (r + s) + 1)}{(1 - p)(r + s)} \right]}{P \left[ F_{(2r, 2[y_1 - r + 1])df} \leq \frac{y_1 p(y_1 - r + 1)}{(1 - p)r} \right]} . \]

confirming that the usual binomial distribution does possess a finite memory. The conditional odds, for a fixed \( y_1 \), for safe asymptomatic symptom are

\[ Odds_{y_2 | y_1} = \frac{\Pr(Y_2 = 0)}{\Pr(Y_2 \geq 1)} = (1 - p)^{y_1} \{1 - (1 - p)^{y_1}\}^{-1} \approx (1 - p)^{y_1} \{1 + (1 - p)^{y_1}\} . \]

The unconditional odds for safe asymptotic symptom are

\[ odds_{y_2} \approx \sum_{y_1=0}^{\infty} Odds_{y_2 | y_1} \Pr[Y_1 = y_1 | \lambda] \approx \sum_{y_1=0}^{\infty} (1 - p)^{y_1} \{1 + (1 - p)^{y_1}\} e^{-\lambda y_1} / y_1 ! \]

\[ \approx e^{-\lambda \hat{\lambda}} \{1 + e^{-p(1-p)\lambda}\} \]
See Figure 4 for the configuration of the odds in asymptotic COVID-19 occurrences in general. Recall that \( S_{Y_2}(1, p \mid y_1) = \Pr(Y_2 \geq 1 \mid y_1) \) is the likelihood for the existence of asymptomatic symptom of COVID-19 in the ship. The hazard in that situation (that is, with \( r = 1 \)) is \( h_{Y_2}(1 \mid y_1, p) = 1 - \frac{\frac{y_1}{2} \hat{p}_{mle}(y_1 - 1)}{2(1 - \hat{p}_{mle})} \), where \( \hat{p}_{mle} = \bar{y}_2 \). A popular statistical concept in the business world (see Khokhlov, 2016 for details), \textit{Tail Value at Risk} (TVaR) is

\[
TVaR_{y_2} = \mathbb{E}[Y_2 \mid Y_2 \geq 1, p, y_1] \approx 1 + \frac{y_1(1 - p)}{p^2 \Pr[F_{(2, 3y_1)_{df}} \leq \frac{y_1^2 p}{(1 - p)}]}.
\]

Similarly, all the Bayesian results for the binomial random variable, \( y_3 \) are easily derivable by interchanging \( \gamma \) and \( \delta \) in all the above expressions. The survival function of the random number, \( Y_3 \) with symptomatic symptoms is

\[
S_{Y_3}(r, p \mid y_1) = \Pr(Y_3 \geq r \mid y_1) = \sum_{i=r}^{\infty} \frac{y_1!}{i!(y_1 - i)!} (1 - p)^i p^{y_1 - i}
\]

\[
= \Pr[F_{(2, 2\{y_1-r+1\})_{df}} \leq \frac{y_1(1 - p)(y_1 - r + 1)}{pr}]; 0 < p < 1.
\]

The hazard rate, \( h(y) \) for the symptomatic sign is

\[
h(y_3) = \frac{\Pr(y_3 \mid p)}{S(y_3 + 1 \mid p)} = \frac{y_1! \left( \frac{1 - p}{p} \right)^{y_1} p^{y_1}}{y_3!(y_1 - y_2)! \Pr[F_{(2, 2\{y_1-r+1\})_{df}} \leq \frac{y_1(1 - p)(y_1 - r + 1)}{pr}]; 0 < p < 1.}
\]

The binomial distribution of those with symptomatic signs has a finite memory.
Pr(Y_3 \geq s | y_3 \geq r) = \frac{\Pr(Y_3 \geq r+s)}{\Pr(Y_3 \geq r)} = \frac{P[F_{(2r,2|y_3-(r+s)+1)]df \leq \frac{y_3(1-p)(y_3-(r+s)+1)}{p(r+s)}}{P[F_{(2r,2|y_3-r+1)]df \leq \frac{y_3(1-p)(y_3-r+1)}{pr}},

confirming that the usual binomial probability trend of those with symptomatic signs does possess a finite memory. The conditional odds, for a fixed y_1, for safe symptomatic symptom are \( Odds_{y_1|y_1} = \frac{\Pr(Y_1 = 0)}{\Pr(Y_3 \geq 1)} = p^{y_1} \{1 - p^{y_1}\}^{-1} \approx p^{y_1} \{1 + p^{y_1}\} \).

The unconditional odds for safe symptomatic symptom are

\[
odds_{y_1} \approx \sum_{y_1=0}^{\infty} Odds_{y_1} \Pr[Y_1 = y_1 | \lambda] \approx \sum_{y_1=0}^{\infty} p^{y_1} \{1 + p^{y_1}\} e^{-\lambda} \lambda^{y_1} / y_1 ! \approx e^{-[1-p]\lambda} \{1 + e^{-[1-p]\lambda}\}.
\]

Figure 3. Distance, \( d(y_1, \lambda) \) in Poisson. Figure 4. Odds for asymptotic

A comparison of odds_{y_2} and odds_{y_1} suggests the odds ratio,

\[
OR_{y_2/y_1} = \frac{odds_{y_2}}{odds_{y_1}} = e^{-[1-2p]\lambda}. \text{ See Figure 5 for the configuration of the isomorphic factor, } e^{-(1-2p)\lambda}. \text{ Recall that } S_{y_1}(1, p | y_1) = \Pr(Y_3 \geq 1 | y_1) \text{ is the chance for the existence of symptomatic symptom of COVID-19. The hazard in that situation (that is, with}

Page 21 of 30
\[ P[F_{(4,2(\gamma_1-1))df}^\text{mle}] \leq \frac{y_1(1-\hat{p}_\text{mle})}{2\hat{p}_\text{mle}}, \] where \( \hat{p}_\text{mle} = \bar{y}_2 \).

**Tail Value at Risk (TVaR)** is

\[ TVaR_{y_i} = E[Y_3|Y_3 \geq 1, p, y_i] \approx 1 + \frac{y_i p}{(1-p)^2 P[F_{(2\gamma_2)df}^\text{mle}] \leq \frac{y_i^2(1-p)}{p}}. \]

Lastly, we develop the Tango index and its significance level over the period.

Tango (1984) proposed an index to detect disease cluster in grouped data. This index received considerable attention in the literature. Following the line of thinking in Tango (1984), we could next assess the MLEs of several entities we estimated and displayed in the Tables 1, 2, and 3. There are three groups of duration.

The group 1 consists of 15\textsuperscript{th} and 16\textsuperscript{th} February 2020. The group 2 encloses data for 17\textsuperscript{th} and 18\textsuperscript{th} February 2020. The group 3 contains data of 19\textsuperscript{th} and 20\textsuperscript{th} February 2020. Two independent contrasts among the three groups are feasible. In an arbitrary style, we select to compare group 1 with group 2 and then group 2 with group 3. For this purpose, we formulate a contrast matrix \( A_{3 \times 3} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \)

where the third column of the matrix needs no explanation. The Tango’s statistic \( T = r' A r \) follows a chi-squared distribution with \( v = 2 \) degrees of freedom (df), where \( r_{3 \times 3} \) is a row vector of the MLE of a chosen entity in of our analytic results in the Table 1 or Table 2 or Table 3. For an example, let \( r' = (68.5, 93.5, 46) \) for the MLE of the COVID-19’s prevalence rate, \( \lambda \) in the groups. Then, the Tango’s test
statistic, \( T = 422.25 \) with \( v = 2 \) df and \( p-value = 2.03975 \times 10^{-92} \). Likewise, the Tango’s test statistic value and its p-value are calculated and displayed in Table 4 for other entities.

Figure 5. The configuration isomorphic factor \( e^{-(1-2p)\lambda} \)

3. Illustration with COVID-19 data in the Diamond Princess Cruises Ship, 2020.

In this section we illustrate all the concepts and expressions of Section 2. Let us consider the COVID-19 data in Table 1 in the Diamond Princess Cruises Ship, 2020. The Diamond Princess is a cruise ship registered in Britain and operated across the earth. During a cruise that began on 20 January 2020, positive cases of COVID-19 linked to the COVID-19 pandemic were confirmed on the ship in February 2020. Over 700 people out of 3,711 became infected (567 out of 2,666 passengers and 145 out of 1,045 crew), and 14 passengers, died. To be specific, on 15th February 2020, 67 people were infected, on 16th February 2020, 70 people were infected, on 17th February 2020, there were 99 COVID-19 cases, on 18th February,
another 88 cases were confirmed. The U.S. government initially asked Japan to keep them on board for 14 days. The U.S. government, however, changed its policy to bring them to the Travis Air Force Base in California and the Base in San Antonio, Texas.

For each specified day in the 1st column in Table 1, the estimate of the COVID-19’s prevalence rate and its variance are calculated using expressions $\hat{\lambda} = \bar{y}_i$ and $\text{Var}(Y_i|\lambda) = s^2_{y_i}$. Both the prevalence and its variability increased and then decreased over the days. However, their correlation, $\hat{\rho}_{y_2,y_3}$ is calculated using the observed numbers on $y_2$ and $y_3$ for each day (see in Table 2) and the estimated correlations had been stable over the days. Substituting $\hat{\lambda} = \bar{y}_i$ and $\text{Var}(Y_i|\lambda) = s^2_{y_i}$ in the expression $\hat{H}_\lambda = \frac{\hat{\lambda}}{\hat{\lambda} + \text{Var}(Y_i|\lambda)}$, we obtained the non-observable heterogeneity and displayed in Table 2. The non-observable Poisson heterogeneity for $y_i$ was high in the beginning day, came down later but only to increase later. Using $\hat{\lambda} = \bar{y}_i$ and $\text{Var}(Y_i|\lambda) = s^2_{y_i}$ in the expression $\hat{H}_{\bar{y}_i} = [1 + \frac{(1 + \bar{\lambda})}{s^2_{\bar{y}_i}}]^{-1}$, we obtained the observable heterogeneity and displayed in the Table 2. The observable Poisson heterogeneity was low in the beginning day, increases and then to decrease. Note in Table 2 that the observable and non-observable Poisson heterogeneities are inversely proportional. In other words, the estimate of the shape and scale parameter in the Bayesian approach are respectively $\hat{\alpha} = \frac{s^2_{\bar{\lambda}}}{\bar{\lambda}^2}$ and $\hat{\beta} = \frac{\bar{\lambda}}{s^2_{\bar{\lambda}}}$ (see their values in Table
2). The shape parameter value decreased consistently over the days. The scale parameter was high to begin with, then increased later. The distance, \(d(y_i, \lambda)\) between the observable and non-observable Poisson mechanism for \(y_i\) is calculated using the expression \(d(y_i, \lambda) = \{\beta(1 - \beta) \pm 1\}(\frac{\alpha}{1 + \beta})^2\) and displayed in Table 2.

Notice that the distance was large to begin with, decreased then but to increase later over the days.

Note that we compute \(\hat{p}_i = \frac{y_2}{y_1}\) for the \(i^{th}\) day. Then, we calculate their average

\[
\bar{p} = \frac{\sum_{i=1}^{2} p_i}{2}
\]

and variance \(\sigma_p^2 = \frac{(p_1 - p_2)^2}{4}\) (see \(\hat{p}\) in Table 1) and it had been steadily increasing over the days since 15\(^{th}\) February 2020. This is something valuable for the medical professionals learning the clinical nature of COVID-19. Using the expression, \(odds_y = e^{-\rho \hat{\lambda}} \{1 + e^{-\rho(1 - \rho)\hat{\lambda}}\}\) in Section 2.2, we calculated the odds for a COVID-19 case to become an asymptomatic type and displayed in Table 2.

### Table 2. Results for Mizumoto et al.’s COVID-19 data in Diamond Princess.

| Date            | \(OR_{\hat{Y}_i}^{\hat{Y}_2}\) | \(H_{\hat{Y}_i}\) | \(\hat{\beta}\) | \(\hat{\alpha}\) | \(d(y_i, \lambda)\) | \(H_{\hat{\lambda}}\) |
|-----------------|---------------------------------|-------------------|-----------------|-----------------|------------------|------------------|
| 15, 16 Feb 2020 | 943.88                          | 0.27              | 15.22           | 1042.72         | 857.81           | 0.93             |
| 17, 18 Feb 2020 | 7.36E+17                        | 0.70              | 1.54            | 144.50          | 18.79            | 0.61             |
| 19, 20 Feb 2020 | 9.69E+11                        | 0.65              | 2.11            | 97.15           | 23.56            | 0.67             |

Table 3. Results for asymptomatic COVID-19 cases in Mizumoto et al. (2020).
Likewise, using the expression \( odds_{y_1} = e^{-\frac{(1-p)\lambda}{1+e^{-\frac{(1-p)\lambda}{p}}}} \), we estimated the odds for a COVID-19 case to become a symptomatic case and displayed in the Table 2. Notice that both odds \( (Odds_{y_1} \text{ and } Odds_{y_2}) \) are low but their odds ratio, \( OR_{\frac{y_2}{y_1}} = e^{-\frac{(1-2p)\lambda}{\lambda}} \), is not negligible but reveals that the situation is favorable to symptomatic rather than asymptomatic. This discovery is feasible because of approach and it is an eye-opening for the medical professionals in their desire to control the spread of the COVID-19. Both the observable, \( \hat{H}_{y_2|y_1} \) and non-observable, \( \hat{H}_{y_1\gamma,\delta} \) binomial heterogeneity (see their values in Table 3) were decreasing for the number, \( y_2 \) of asymptomatic COVID-19 cases. The distance, \( d(y_2, p) \) between the observable and non-observable for asymptomatic cases was moderate in the beginning, then increased, and then decreased over the days (see their values in Table 3). However, the distance, \( d(Y_2, Y_1) \) between the observable,
$y_2$ of the asymptomatic cases and the observable, $y_3$ of the symptomatic cases was narrow, then wider, and then moderate over the days (see their values in Table 3).

Table 4. Tango’s test statistic and its p-value for several entities

| Tango statistic | $\hat{H}_{y_1}$ | $H_\lambda$ | $\bar{p}$ | $\hat{H}_{y_1,y,\delta}$ | $\hat{H}_{y_2|y_1}$ | $d(y_1,\hat{\lambda})$ | $d(y_2,p)$ | $d(Y_2,Y_3)$ |
|-----------------|-----------------|-------------|-----------|-----------------|------------------|--------------------|-------------|-------------|
| T with 2 df     | 0.25            | 0.36        | 0.41      | 0.77            | 0.51             | 699420             | 260.66      | 751.20      |
| p-value         | 0.87            | 0.83        | 0.81      | 0.67            | 0.777            | 0.0E100            | 2.4E-57     | 164         |

For a COVID-19 case to become a symptomatic type, the chance is moderate to less and more over the days (see $1 - \bar{p}$ in Table 3). The estimate of the shape and scale parameter happened to be $\hat{\gamma}$ and $\hat{\delta}$ respectively (see their values in Table 3). Both the shape parameter and the scale parameter values decreased drastically over the days. From the p-values in the Table 4, we infer that the prevalence rate, $\hat{\lambda}$, the distances, $d(y_1,\hat{\lambda})$, $d(y_2,p)$ and $d(Y_2,Y_3)$ do differ significantly over the three groups of days. The chance for a COVID-19 to become an asymptomatic type does not differ significantly across the three groups. On the contrary, the non-observable heterogeneities $H_\lambda$ of the Poisson random number, $y_1$ and $\hat{H}_{y_1,y,\delta}$ of the binomial random number, $y_2$ are not significant. Likewise, the observable heterogeneities $\hat{H}_{y_1}$ of the Poisson random number, $y_1$ and $\hat{H}_{y_2|y_1}$ of the binomial random number, $y_2$ for a given $y_1$ are not significant.
4. Conclusion with comments.

The risk of contracting COVID-19 in cruise is more than in a community as the social distancing is weakly implemented and the breathing air is tightly internalized. More nations are afraid to let the voyagers come inside at the seaports. Not even the ships are permitted to dock at the port to avoid mitigation by the communities. The scenario seems to be ant humanistic. The crews are known to have committed suicide in the ship itself. The medical doctors and/or pharmaceutical service are strained for the infected and COVID-19 free voyagers. Lack of clear symptoms among those that are infected adds to difficulties in managing the COVID-19 crisis in any ship. Most importantly, how do we dispose the COVID-19 death bodies, away from further infectivity.

In the midst of uncertainties about the root cause and/or the appearance of any symptoms, the best modelers can do (as it is done in this article) is to devise a methodology to address the observable as well as non-observable heterogeneity, estimate the proportion of COVID-19 cases to be asymptomatic, estimate the odds of becoming symptomatic, and also the odds ratio for asymptomatic in comparison to those symptomatic among the COVID-19 cases. Some of these are non-trivial to the professional experts dealing with the reduced spread of COVID-19 if not its total elimination. Much of COVID-19 is mystic.

References

1. Blumenfeld, D. (2010) *Operations Research Calculations Handbook*, CRC Press, Boca Raton, Florida.
2. Ecochard, J. (2006) Heterogeneity in fecundability studies: issues and modeling, *Statistical Methods in Medical Research, 15, 141-160.*

3. Elston, R., Olson, J., Palmer, L. (2002) *Biostatistical Genetics and Genetic Epidemiology,* Wiley Press, Baffins Lane, Chichester, West Sussex, UK.

4. Hope, T. and Norris, P. (2013) Heterogeneity in the frequency distribution of crime victimization, *Journal of Quantitative Criminology, 29 (4), 543-576.*

5. Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., Glasziou, P. F. (2018) *Decision making in health and medicine: Integrating evidence and values,* Cambridge University Press, Cambridge, UK.

6. Khokhlov, V. (2016) Conditional value-at-risk for elliptical distributions, *Evropský časopis Ekonomiky a Managementu.* 2 (6), 70–79.

7. Mizumoto, K. and Chowell, G. (2020) Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020, *Infectious Disease Modelling 5 (2020) 264-270.*

8. Mizumoto, K., Kagaya, K., Zarebski, A., Chowell, G. (2020) Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19), cases on board the Diamond Princess cruise ship, Yokohama, 2020. *Euro Surveill.* 25(10), 1560-7917.

9. Rajan, C. and Shanmugam, R. (2020) *Discrete Distributions in Engineering and the Applied Sciences,* *Synthesis Lectures on Mathematics and Statistics,*
10. Ross, Sheldon. (2002) *A First Course in Probability*, sixth edition, Prentice Hall, Upper Saddle River, New Jersey 07458, page 392.

11. Shanmugam, R. (2020) Probabilistic Patterns Among Coronavirus Confirmed, Cured and Deaths in Thirty-two India’s States/Territories, accepted to appear in *International Journal of Ecological Economics and Statistics*, page numbers are pending.

12. Shanmugam, R. and Radhakrishnan, R. (2011) Incidence jump rate reveals over/under dispersion in count data, *International Journal of Data Analysis, and Information Systems*, 3, 1, 1-8.

13. Spreeuw, J. (1999) Heterogeneity in hazard rates in insurance, Tinbergen Institute of Research Series, 210, *Ph. D. thesis*, University of Amsterdam, Amsterdam.

14. Stuart, A. and Ord, K. (2015) *Kendall’s Advanced Theory of Statistics, Volume 1*. Oxford University Press, London.

15. Tango, T. (1984) The detection of disease clustering in time, *Biometrics*, 40, 15-26.