Scientific Research Publishing

Advances in Pure Mathematics, 2019, 9, 205-227
http://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Deconvolution of the Error Associated with Random Sampling

Peter L. Irwin*, Yiping He, Chin-Yi Chen

Molecular Characterization of Foodborne Pathogens, United States Department of Agriculture,
Wyndmoor, PA, USA
Email: *peter.irwin@ars.usda.gov

How to cite this paper: Irwin, P.L., He, Y. and Chen, C.-Y. (2019) Deconvolution of the Error Associated with Random Sampling. *Advances in Pure Mathematics*, **9**, 205-227.
https://doi.org/10.4236/apm.2019.93010

Received: February 11, 2019
Accepted: March 26, 2019
Published: March 29, 2019

Abstract

In this work empirical models describing sampling error (Δ) are reported based upon analytical findings elicited from 3 common probability density functions (*PDF*): the Gaussian, representing any real-valued, randomly changing variable x of mean μ and standard deviation σ; the Poisson, representing counting data: *i.e.*, any integral-valued entity's count of x (cells, clumps of cells or colony forming units, molecules, mutations, etc.) per tested volume, area, length of time, etc. with population mean of μ and $\sigma = \sqrt{\mu}$; binomial data representing the number of successful occurrences of something (x^{+}) out of n observations or sub-samplings. These data were generated in such a way as to simulate what should be observed in practice but avoid other forms of experimental error. Based upon analyses of 10^4 Δ measurements, we show that the average Δ ($\bar{\Delta}$) is proportional to $\sigma \cdot \sqrt[-2]{n} \cdot \mu^{-1}$ ($\sigma_{\bar{x}} \cdot \mu^{-1}$; Gaussian) or $\sqrt[-2]{n \cdot \mu}$ (Poisson & binomial). The average proportionality constants associated with these disparate populations were also nearly identical ($\bar{A} = 0.783 \pm 0.0470$; ±*s*). However, since $\sqrt{\mu} = \sigma$ for any Poisson process, $\sqrt[-2]{n \cdot \mu} = \sigma_{\bar{x}} \cdot \mu^{-1}$. In a similar vein, we have empirically demonstrated that binomial-associated $\bar{\Delta}$ were also proportional to $\sigma_{\bar{x}} \cdot \mu^{-1}$. Furthermore, we established that, when all $\bar{\Delta}$ were plotted against either $\sqrt[-2]{n \cdot \mu}$ or $\sigma_{\bar{x}} \cdot \mu^{-1}$, there was only one relationship with a slope = A (0.767 ± 0.0990) and a near-zero intercept. This latter finding also argues that all $\bar{\Delta}$, regardless of parent *PDF*, are proportional to $\sigma_{\bar{x}} \cdot \mu^{-1}$ which is the coefficient of variation for a population of sample means ($C_V\left[\bar{x}\right]$). Lastly, we establish that the proportionality constant A is equivalent to the coefficient of variation associated with Δ ($C_V\left[\Delta_j\right]$) measurement and, therefore, $\bar{\Delta} = C_V\left[\Delta_j\right] \cdot C_V\left[\bar{x}\right]$. These results are noteworthy inasmuch as they provide a straightforward empirical link between stochastic sampling error and the aforementioned C_Vs. Finally, we demonstrate that all attendant empirical measures of Δ are reasonably small (e.g., $s_{\bar{x}} \cdot \bar{x}^{-1} \sim 4\%$) when an environmental microbiome was well-sampled: n = 16 - 18 observations with $\mu \sim 3$ isolates per observation. These colony counting results were supported by the fact that the two major isolates' relative abundance was reproducible in the four most probable composition observations from one common population.

Keywords

Stochastic Sampling Error, Modeling, Most Probable Composition, Quantitative Metagenomics, Food-Borne Bacteria

1. Introduction

There are various analytical procedures for enumerating organisms in environmental samples which diverge in their experimental approach yet are mathematically inter-related. Thus, if V represents the sample volume and V_e the volume occupied by a test entity of interest (e.g., colony forming units or *CFU*s), the probability that *one* particular V_e will *not* contain this entity at concentration δ [1] is

$$\left(\frac{V/V_e - V\cdot\delta}{V/V_e}\right) = \left(1 - V_e\cdot\delta\right);$$

i.e., V/V_e—maximum possible number of entities in V and $V\cdot\delta$~the actual number of objects present.

Assuming that many V_e aliquots have been combined to generate V, the probability that *no* organism will be contained in V is [1]

$$P^- = \left[1 - V_e\cdot\delta\right]^{\frac{V}{V_e}}$$

therefore

$$\ln\left[P^-\right] = \frac{V}{V_e}\cdot\ln\left[1 - V_e\cdot\delta\right].$$

Since

$$\ln\left[1-\psi\right] \sim -\psi - \frac{\psi^2}{2} - \frac{\psi^3}{3} - \frac{\psi^4}{4} - \cdots$$

then, if $\psi = V_e\cdot\delta$,

$$\begin{aligned}\ln\left[P^-\right] &\sim \frac{V}{V_e}\left(-V_e\cdot\delta - \frac{V_e^2\cdot\delta^2}{2} - \frac{V_e^3\cdot\delta^3}{3} - \frac{V_e^4\cdot\delta^4}{4} - \cdots\right) \\ &\sim -V\cdot\delta\left(1 + \frac{V_e\cdot\delta}{2} + \frac{V_e^2\cdot\delta^2}{3} + \frac{V_e^3\cdot\delta^3}{4} + \cdots\right).\end{aligned}$$

For $V_e \to 0$ (e.g., *E. coli* [2] has a $V_e \sim 0.6\ \mu\text{m}^3 \sim 6\times10^{-13}\,\text{mL}$),

$$\ln\left[P^{-}\right] \sim -V \cdot \delta$$

$$P^{-} = \exp\left[-V \cdot \delta\right] = \exp\left[-\mu\right]$$

therefore

$$P^{+} = 1 - P^{-} = 1 - \exp\left[-V \cdot \delta\right] = 1 - \exp\left[-\mu\right]. \tag{1}$$

In certain circumstances it is *only* possible to determine an organism's δ by diluting the sample to such an extent that only a fraction of the n "technical" replicates tested are positive (x^{+}) for the presence of the entity, or microbe, in question [3] [4]. This technique is referred to as the "dilution method" [1] since it involves diluting a test sample's content to extinction ($\delta \to 0$). This enumeration protocol is also known as the most probable number (*MPN*) method and entails sampling from a liquid source, making serial dilutions from this, distributing an aliquot of each of these dilutions into separate receptacles, incubating these under suitable growth conditions, and observing if any growth has occurred based upon some organism-specific detection method [5] [6]. The *MPN* enumeration procedure is particularly useful when sampling from environmental sources, such as foods, since damaged cells frequently recover in liquid media [7].

For example, were one to obtain a food sample containing ~14 *CFU* of a particular organism per 50 g, the cells would typically be washed from the food matrix, concentrated to a few mL (e.g., via centrifugation), and brought up to some appropriate volume (say 40 mL = V_{sample}) with media [5]. From this, eight 4 mL (V) samples could be randomly selected and distributed into 8 separate receptacles (n = 8 with a dilution factor of 1; *i.e.*, undiluted). Of the remaining 8 mL, 4 could be further diluted with 36 mL (40 mL total) liquid media, mixed and distributed into another set of 8 containers. This set of dilutions has a dilution factor of 0.1 relative to the original. With the remaining 8 mL from the 0.1 dilution, 4 mL could be diluted again with 36 mL media, mixed and distributed into yet another eight 4 mL replicates (dilution factor = 0.01). After incubation the most likely number (Equation (2), below) of positive occurrences (e.g., presence of a specific gene [5]) observed would be x^{+} = 6, 1, and 0 (out of n = 8 observations per dilution) for dilution factors of 1, 0.1, and 0.01, respectively, and the calculated *MPN* (±*s*) per 50 g sample would = 13.8 ± 5.56. Note the relatively large error term. For a 4-fold proportional (200 g, 160 mL V_{sample}) experiment with n = 32, the calculated *MPN* is 13.8 ± 2.78 per 50 g sample.

For *MPN*-based organism detection and subsequent enumeration, the number of positive occurrences of growth in any j^{th} experiment out of n observations = $x_j^{+} = \sum_{i=1}^{n} \theta_{ij}$ (θ = either 1 [presence] or 0 [absence]) can be estimated as

$$x^{+} \sim n \cdot P^{+} = n\left(1 - \exp\left[-V \cdot \delta\right]\right) \tag{2}$$

whereupon x^{+} is integral (=ROUND($n \cdot P^{+}$, 0) in Excel). The probability of observing x^{+} successes out of n Bernoulli trials [8] each of volume V from a population of δ entities per V is

$$P_b = \frac{n!}{x^+!\left(n-x^+\right)!}\left(P^-\right)^{n-x^+}\left(P^+\right)^{x^+}$$

which is also known as the binomial *PDF*. Since $n \cdot P^+$ = the population average (real) [9] number of positive responses out of *n* tests (μ^+), the above can be also written as

$$P_b = \frac{n!}{x^+!\left(n-x^+\right)!}\left(1-\frac{\mu^+}{n}\right)^{n-x^+}\left(\frac{\mu^+}{n}\right)^{x^+}. \tag{3}$$

The multiple dilution *MPN* calculation itself is determined by finding the value of δ at the maximum in the product of the $P_b s$ from all ℓ^{th} dilutions ($\prod_\ell P_{b,\ell}$) and is easily achieved by adding the scaled sum of all dilutions' $\partial_\delta P_b \div P_b$ values to an initial guess for δ (*i.e.*, $\delta_{m+1} = \delta_m + \lambda_m \times \sum_\ell \left\{\partial_\delta P_{b,\ell} \div P_{b,\ell}\right\}_m$ $= \delta_m + \lambda_m \times \sum_\ell \left\{\left(x_\ell^+ - n + \left(x_\ell^+ \div \left(\exp\left[V \cdot \delta_m \cdot 0.1^\ell\right] - 1\right)\right)\right) \cdot V \cdot 0.1^\ell\right\}$ for any particular ℓ^{th} one-to-ten dilution and *m* iterations; λ is a monotonically changing, with *m*, scaling function) then solving for the *MPN* recursively [1] [4] [5] [10] which minimizes the summation.

At the limit *n* → ∞, Equation (3) simplifies to what is known as the Poisson *PDF*

$$P_P = \frac{\mu^x \exp[-\mu]}{x!}. \tag{4}$$

Under these circumstances, *x* is the observed and μ is the population average number of counts in/on the tested volume, surface, chosen time period, etc. This *PDF* is applicable to all analytical systems involving, essentially, the counting of objects. However this *PDF* is applied, the most conspicuous aspect [11] [12] of any Poisson process is that the *variance* (σ^2 or second moment)

$$\sigma^2 = \sum_{x=0}^{\infty}(x-\mu)^2 P_P = \mu$$

equals the *population mean* (μ or first moment)

$$\mu = \sum_{x=0}^{\infty} x \cdot P_P .$$

The last probability density function utilized in this stochastic sampling exercise is also related to P_b, Equation (3). This is the Gaussian *PDF* which we use to quantitatively examine the effects of *n* and σ (fixed μ) on the variability of sample means ($\bar{x}$) which have been created by randomly sampling from a population of real-valued variables (*x*; e.g., doubling time [13]) which are normally distributed as

$$P_G = \frac{\text{Area}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; \tag{5}$$

in this relationship the *Area* term (~ $\Delta x \cdot \sum_{k=1}^{K} f_k$; for large *K*) is the approximate area under the fitting function *f* (frequently taken to be 1 since Δx is often = 1 and $\sum f$ is always ~1). There are several derivations of P_G but none are as persuasive as the fact that this *PDF* is simple and has been experimentally shown to be the most likely probability distribution associated with most experimental observations [9] [12].

The original purpose of our sampling-related investigations [7] was to estimate a nominal value for *n* needed to achieve accurate most probable foodborne bacterial isolate enumeration, combined with 16S *rDNA*-based identification, for quantitative metagenomic purposes. The relationships were developed by examining the results of 6 × 6 colony counting (Poisson *PDF*) of highly diluted bacteria [14] [15] as a function of *n* and μ as well as by generating counts (*x*) derived from P_P to simulate what occurred in the lab [15] [16] but which avoided other forms of experimentally based error [5]. We were able to establish that $n_{\min} = n_{\mu\to1} \div \sqrt[3]{\mu}$ where $n_{\mu\to1}$ is the number of observations necessary to accurately enumerate a population average of 1 count per volume tested. Based mainly on colony counting experience we estimate $n_{\mu\to1}$ is somewhere in the range *n* ~ 20 - 30 observations.

Herein we model stochastic sampling errors associated with *all* the aforementioned *PDF*s and empirically demonstrate that the resultant mathematical models are, in part, a consequence of the "central limit theorem" [17] (*CLT*). In general, the *CLT* states that a distribution of sample means ($\bar{x}$), regardless of parent *PDF*, approaches a normal distribution analytically equivalent to P_G, Equation (5), with $x = \bar{x}$, $\mu = \mu_{\bar{x}}$, and with the σ^2 term = $\sigma_{\bar{x}}^2$ (= $\sigma^2 \div n$) as the number of separate *n*-samplings increases. We also have elaborated on empirical findings developed previously [5] [15] [16] for predicting errors associated with the random sampling of microorganisms as well as comparing the internal variations associated with the three different sampling error data types derived from the Gaussian, binomial (*MPN*), and Poisson relationships. Thus, new results have been created using the aforementioned probability distributions, Equations (2), (4), and (5), and have been highly replicated since each "experiment", comprising *n* (= 3, 6, 9, 12, or 24) observations, were repeated 100 times.

2. Materials and Methods

2.1. Poisson-Based Data: Equation (4), Figure 1

All counting data were created by multiplying Equation (4) by 360 in order to produce a large number of integral-valued repeats (=ROUND ($360 \cdot P_P$, 0)) for any particular count *x*: e.g., for $\mu = 1$ particle per test volume, area, length of time, etc., there would be, most probably, 132 repeats of *x* = 0, 132 repeats of *x* = 1, 66 repeats of *x* = 2, 22 repeats of *x* = 3, 6 repeats of *x* = 4 and 1 repeat of *x* = 5 entities per test. From this pool of 360 counts for each *μ*, an *n* number of *x* values were randomly selected based upon random number tables created with *Mathematica*.

$$\text{Table}\left[i = \text{Random}\left[\text{Integer}, \{1{,}360\}\right], \{i{,}n\}\right] \tag{6}$$

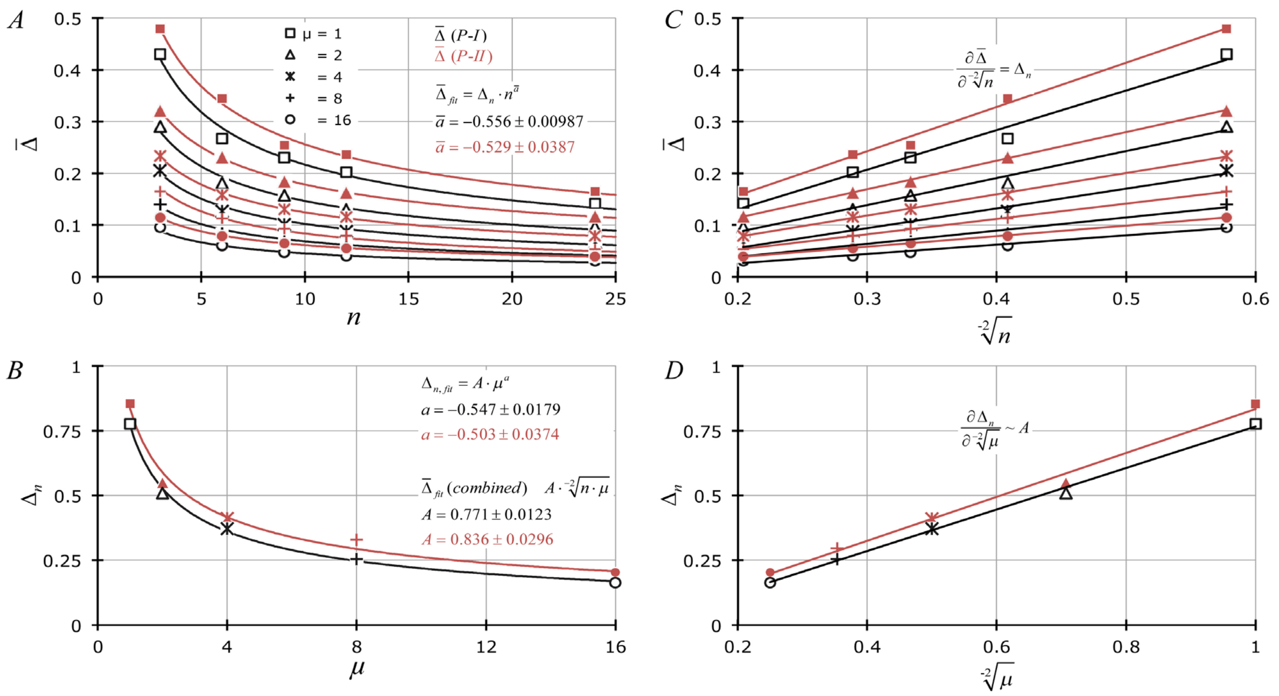


Figure 1. (A) relationship of average Δ_j ($\bar{\Delta}$) for Poisson-based data using Equation (7) (*P-I*: black symbols and curves) or Equation (8) (*P-II*: red symbols and curves) as a function of n (= 3, 6, 9, 12, 24) and various values for μ (= 1, 2, 4, 8, 16). Gauss-Newton least squares minimization-based curve-fitting [18] of data was performed [19] to fit to the equation $\bar{\Delta} = \Delta_n \cdot n^{\bar{a}}$ (averages for a are provided ± s; averaged across 5× μ). (B) Non-linear relationship of individual Δ_n values from (A) for *P-I*- and *II*-based data as a function of μ whereupon curve-fitting of data was also performed using the algebraic form $\Delta_n = A \cdot \mu^a$ (values for A and a are provided ± *ASE*). (C) and (D) Present linearized forms ($X = \sqrt[-2]{n}$ in (C) and $X = \sqrt[-2]{\mu}$ in (D)) of data reported in Figure 1(A) and Figure 1(B) based upon all values of a = −1/2. Slopes of the lines in Figure 1(C) and Figure 1(D) are equivalent to Δ_n and A, respectively.

which generates n random numbers between 1 and 360. Thus, 100 such random number sets were utilized for the twenty-five n (= 3, 6, 9, 12, 24) × μ (= 1, 2, 4, 8, 16) combinations. Briefly, each procedure involved arranging the aforementioned 360 x values (one set for each μ) in one column of a spreadsheet followed by filling in n adjacent columns with formulae which refer to the calculated x values but where each row's reference number was taken from the *Mathematica*-generated random number, Equation (6), next in sequence. *MPN*- and Gaussian-based data arrays were treated in an identical fashion. The formula (*P-I*: normalized deviations of s_j from $\sigma = \sqrt{\mu}$) for calculating our empirical measure of Poisson stochastic sampling error (Δ) was

$$\Delta_j = \frac{\left|\sqrt{\mu} - s_j\right|}{\mu} \tag{7}$$

whereupon the s_j term is the experimental standard deviation ($\sqrt{(n-1)^{-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$ or "=STDEV.S (x_{ij}-array)" in Excel) for each j^{th} ($j = 1, 2, \cdots, J$; J = 100) experiment and i^{th} ($i = 1, 2, \cdots, n$) x. The average across 100× experiments, regardless of formulation, were symbolized as $\bar{\Delta}$ (=

$J^{-1}\cdot\sum_{j=1}^{J}\Delta_j$ or “=AVERAGE (Δ_j-array)”). A second form for the Poisson-based measure of Δ was also calculated (*P-II*: normalized deviations of $\bar{x}_j$ from known *μ*) from these same data

$$\Delta_j = \frac{\left|\mu - \bar{x}_j\right|}{\mu}. \tag{8}$$

Here the $\bar{x}_j$ is the observed arithmetic mean for each j^{th} counting experiment.

2.2. MPN Experiments: Equation (1), Figure 2

All *MPN* data were created by multiplying Equation (1) by 360 to produce the number (“=ROUND ($360\cdot P^{+}$, 0)”) of positive responses (*θ* = 1) for any particular level of $V\cdot\delta$ (=*μ*); e.g., for $\mu = 0.1$ entity per volume tested there would be 34 repeats of *θ* = 1 and 326 repeats of *θ* = 0. From such a column of 360 *θ* values (one column for each *μ*), *n* were randomly selected based upon *Mathematica* tables, Equation (6), and treated similar to the Poisson data above. Thus, for each combination of *n* (= 3, 6, 9, 12, or 24) × *μ* (= 0.1, 0.2, 0.4, 0.8, 1.6), 100

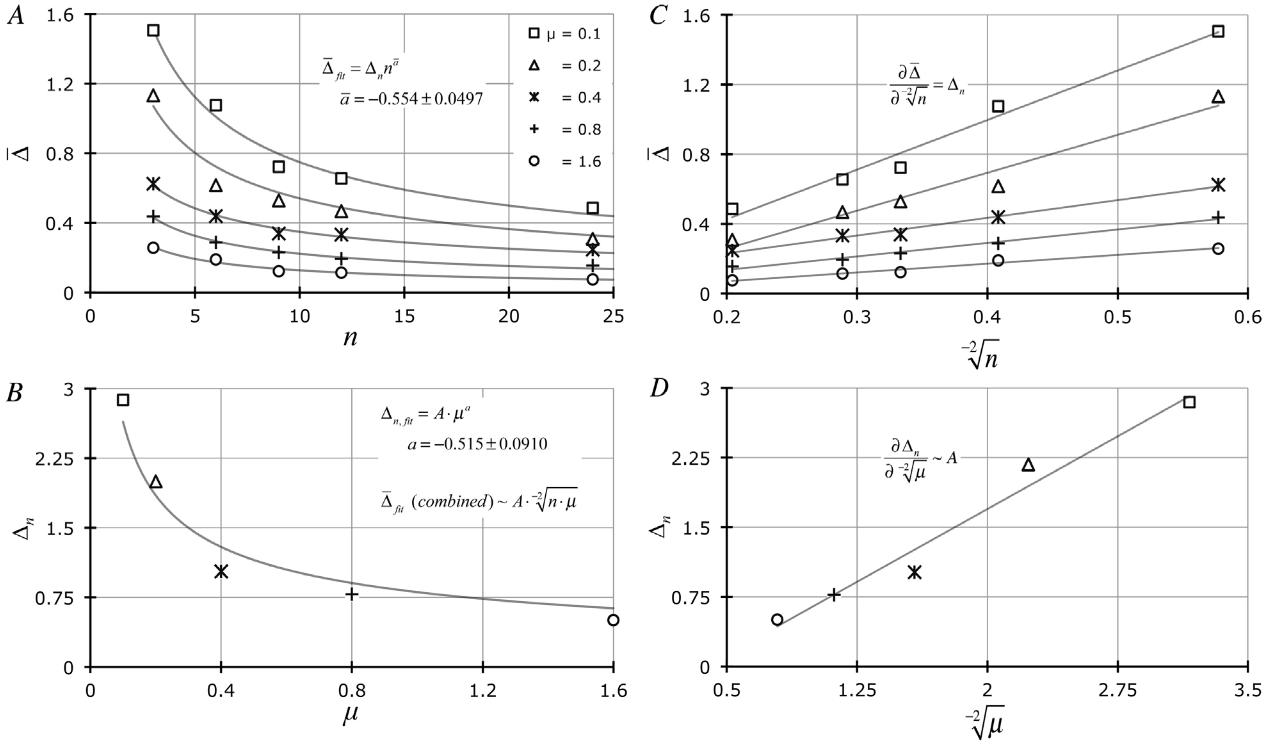


Figure 2. (A) Relationship of average Δ_j ($\bar{\Delta}$) for *MPN*-based data using Equation (9) as a function of *n* (= 3, 6, 9, 12, or 24) and variable *μ* (= 0.1, 0.2, 0.4, 0.8, 1.6). Gauss-Newton least squares minimization-based curve-fitting [18] of data was performed [19] to fit the algebraic form $\bar{\Delta} = \Delta_n\cdot n^{\bar{a}}$ (averages for *a* are provided ± *s*; averaged across 5 × *μ*) to these results. (B) Relationship of individual Δ_n values from (A) for *MPN*-based data as a function of *μ* where curve-fitting of data was performed also to the algebraic form $\Delta_n = A\cdot\mu^{a}$ (values for *A* and *a* are provided ± *ASE*). (C) and (D) Represent linearized forms ($X = \sqrt[-2]{n}$ in (C) and $X = \sqrt[-2]{\mu}$ in (D) of data reported in Figure 2(A) and Figure 2(B) based upon the assumption that *a* = −1/2. Slopes of the lines in Figure 2(C) and Figure 2(D) are equivalent to Δ_n and *A*, respectively.

random n-selections were performed. The formula for calculating our empirical measure of *MPN* sampling error was

$$\Delta_j = \frac{\left| n \cdot P^+ - \sum_{i=1}^{n} \theta_{ij} \right|}{n \cdot P^+} = \frac{\left| \mu^+ - x_j^+ \right|}{\mu^+}; \tag{9}$$

where $\theta =$ either a "1" (a positive occurrence) or a "0" (a negative occurrence). As before, the average Δ_j across $J = 100$ experiments (each of n observations) $= \bar{\Delta}$. The *MPN* value for $\bar{x}_j^+ = \ln\left[n \div \left(n - x_j^+ \right) \right]$ and provides the average *MPN* or *CFU* per sample; a rearrangement of Equation (2).

2.3. Gaussian-Based Data: Equation (5), Figure 3

All Gaussian *PDF* data were produced by multiplying Equation (5) ($\Delta x = 1$) by 360 producing an integral number of observations ("=ROUND ($360 \cdot P_G$, 0)") for each value of x as a function of μ (fixed at 20) and σ (= 1, 1.5, 2, 3, 4). For instance, for $\sigma = 1$ there would be 2 repeats of $x = 17$, 19 repeats of $x = 18$, 87 repeats of $x = 19$, 144 repeats of $x = 20$, 87 repeats of $x = 21$, 19 repeats of $x = 22$, and 2 repeats of $x = 23$. From this column of 360 values of x, n (= 3, 6, 9, 12, or

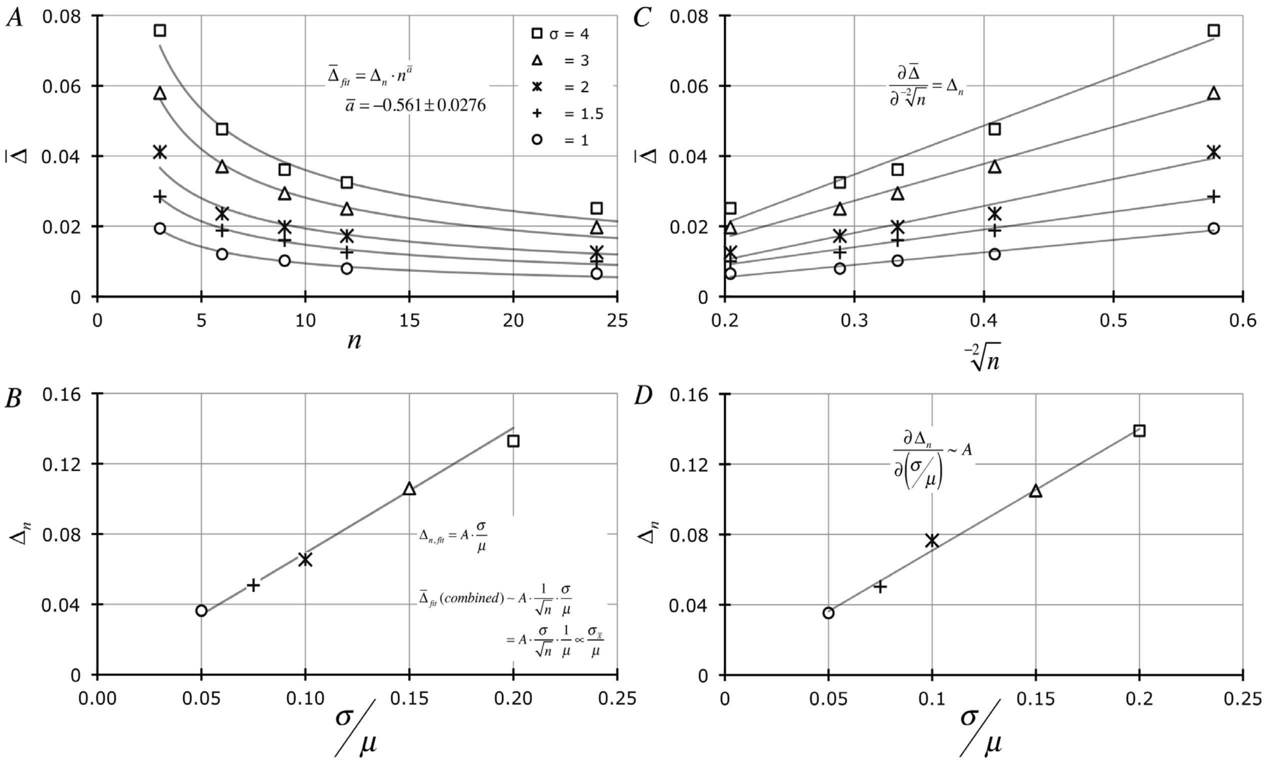


Figure 3. (A) Relationship of average Δ_j ($\bar{\Delta}$) for Gaussian-based data using Equation (10) as a function of n with variable σ (=1, 1.5, 2, 3, 4; μ = 20). Gauss-Newton least squares minimization-based curve-fitting [18] of data was performed [19] to fit the equation $\bar{\Delta} = \Delta_n \cdot n^{\bar{a}}$ (averages for a are provide ± s; averaged across 5× σ) to these results. (B) and (D) Relationship of individual Δ_n values from (A) and (C) for Gaussian-based data as a function of μ-normalized standard deviations ($X = \sigma \div \mu$). Linear regression-based fitting of data was performed to the algebraic form $A \cdot \sigma \div \mu$. Figure 3(C): linearized forms ($X = \sqrt[-2]{n}$) of data reported in Figure 3(A) based on $a = -1/2$. Slopes of the lines in Figure 3(C) are equivalent to Δ_n and plotted in Figure 3(D).

24) were randomly selected based upon Equation (6) and treated identically to the Poisson and *MPN* data sets. Thus, for each combination of $n \times \sigma$ 100× n-based selections were performed. The formula for calculating our empirical measure of Gaussian sampling error, similar to Equation (7), was

$$\Delta_j = \frac{\left|\sigma - s_j\right|}{\mu}. \tag{10}$$

As usual, the average Δ_j across J = 100 such sets of experiments each of n observations = $\bar{\Delta}$.

2.4. Other Calculations

All curve-fitting was based upon a modified Gauss-Newton algorithm by least squares [18] minimization performed on a Microsoft Excel spreadsheet: [19] some of these results were fit to the algebraic form $f\left[X\right] = \text{constant} \cdot X^a$. However, certain *MPN* data ($\bar{x}^+$ and x^+) were also fit to a Gaussian (Equation (5): $P_G\left[\bar{x}^+\right]$ or $P_G\left[x^+\right]$) with Δx used as one of the parameters to be iteratively resolved (*i.e.*, deconvolved). Where appropriate, confidence limits (*CL*) have been calculated using an approach applicable to any hypothetical fitting function $f_k = f\left[X_k;\pi_p\right]$: $k = 1,2,\cdots,K$ rows of the observed *X*-*Y* data sets with up to *P* (typically ≤ 3) fitting parameters π_p ($p = 1,2,\cdots,P$). In this procedure we use the propagation of error method [9] [20] for estimating the standard error associated with each f_k (s_{f_k}; illustrated below for *P* = 2 fitting parameters) data point

$$CL = t \cdot s_{f_k} = t\sqrt{s_{\pi_1}^2\left[\partial_{\pi_1} f_k\right]^2 + s_{\pi_2}^2\left[\partial_{\pi_2} f_k\right]^2 + 2 \cdot s_{\pi_1\pi_2}^2 \cdot \partial_{\pi_1} f_k \cdot \partial_{\pi_2} f_k}$$

where, for any particular fitting parameter ω, $s_\omega = \sqrt{s_Y^2 \cdot \left[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right]_{\omega\omega}^{-1}}$ = "asymptotic standard error" [19] (*ASE*; s_Y^2 = residual sum of squares ÷ [*K* − *P*]), and the $\partial_{\pi_p} f_k$ terms symbolize $\partial f_k / \partial \pi_p$. The above equation simplifies to

$$CL = t_{0.01} \cdot s_{f_k} = t_{0.01}\sqrt{s_Y^2\left(\boldsymbol{Z}_k\left[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right]^{-1}\boldsymbol{Z}_k^{\mathrm{T}}\right)}.$$

In all the above relationships *Z* is the partial first derivative matrix of f_k with respect to the parameters π_1 and π_2 (*i.e.*, a 2-parameter fit) such that

$$\boldsymbol{Z} = \begin{bmatrix} \partial_{\pi_1} f_1 & \partial_{\pi_2} f_1 \\ \partial_{\pi_1} f_2 & \partial_{\pi_2} f_2 \\ \vdots & \vdots \\ \partial_{\pi_1} f_K & \partial_{\pi_2} f_K \end{bmatrix},$$

$\boldsymbol{Z}^{\mathrm{T}}$ is the transpose of *Z*, $\boldsymbol{Z}_k = \left[\partial_{\pi_1} f_k \quad \partial_{\pi_2} f_k\right]$ (*K* row vectors), and $s_Y^2 \cdot \left[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right]^{-1}$ is the variance-covariance matrix [21]. *CL* were not used for all results since they might have muddled analytical aspects of the compositions.

2.5. Microbiome Sampling Data

For the food microbiome sampling experiment ~25 g of commercial, pre-thawed

(~15 min at room temperature), frozen vegetables were washed with a volume of phosphate buffered saline (*PBS*; 10 mM Na_2HPO_4 + 2 mM NaH_2PO_4 + 137 mM NaCl; pH 7.4 ± 0.2; Boston BioProducts, 159 Chestnut Street, Ashland, MA 01721) equivalent to double the mass of the sample. In order to assist in the detachment of plant tissue-bound cells, 0.075% [w/v] Tween-20 (Sigma-Aldrich, 3050 Spruce St., St. Louis, MO 63103) was added to the *PBS* and filter sterilized. All washing was performed in sanitized plastic zip-lock bags wherein the formerly frozen vegetables and buffer wash were gently agitated at 80 rpm for approximately 20 min and immediately passed through a 40 μm nylon filter (BD Falcon; Becton Dickinson Biosciences, Bedford, MA) to remove large particles.

Directly sampled washes (5 mL Control = Observation I [cultured at 30˚C] and III [cultured at 37˚C]) as well as hollow fiber microfilter-concentrated (each 5 mL sample was diluted to ~100 mL *PBS* + Tween, concentrated, then washed with another 100 mL buffer, and eluted with ~5 mLs *PBS* + Tween = Observation II [cultured at 30˚C] and IV [cultured at 37˚C]) samples were collected and enumerated using the 6 × 6 drop plate method [14] but using 1:2 serial dilutions for colony selection on Brain Heart Infusion agar (*BHI* + 2% [w/v] agar). Briefly, this drop plate method involved loading 400 μL of each wash (either control or concentrated samples brought back to the control sample's original volume = 5 mL) filtrate into the first well (row A) of a 96-well microtiter plate. Two-fold serial dilutions were made by transferring 200 μL (multichannel pipette, Rainin, Emeryville, CA) from the first row (row A; dilution 0) into 200 μL of diluent (*PBS*) in the 2nd row (row B; dilution 1), mixing 10 times while continuously stirring, and repeating the process until five 1:2 dilutions were produced; pipette tips were changed between dilutions. Based on a previous analysis of 6 × 6 drop plate sampling error [15], we sampled $n = 16 - 18$ seven μL volumes from each of the 6 dilutions (dilutions 0 - 5; overall dilution factors of $0.5^0 = 1$ to $0.5^5 = 0.03125$) and drop-plated these onto *BHI* agar media using a multichannel pipette. After plating, the droplets were allowed to dry, inverted and then incubated at two temperatures (either 30˚C or 37˚C; 3 plates for each temperature and treatment combination). Colonies were counted after 16 - 24 hours. Colony collection for our 16S *rDNA* bacterial identification protocol [7] involved selecting all colonies from dilution 2 ($0.5^2 = 0.25$ dilution; $\bar{x} = 2.79 \pm 1.52$ colonies per drop; ± s; the fact that $\sqrt{\bar{x}} = 1.67 \sim s$ might argue for an appropriately sampled population).

Each colony (n total) was carefully removed from the agar plate's surface using a Rainin L20 tip, dispersed into 200 μL BHI in a 96-well plate and incubated at 30˚C for 16 - 24 hours. These cultures were restreaked onto solid media and incubated at 30˚C overnight. One colony from each of the original n plates was selected, suspended into 25 μL of Ultra PrepMan (Applied Biosystems, Foster City, CA) in a *PCR* tube and heated in a thermocycler at 99˚C for 15 min. Upon cooling, samples were centrifuged 10 min. to separate the *DNA* solution from the cell debris. A sample of supernatant was transferred to a new tube for the

DNA amplification step (end-point *PCR*). Once the 16S *rRNA* "gene" amplification, sequencing reactions (*EubA* and *EubB* primers) and Sanger sequencing were performed, *DNA* sequences were edited, and contigs assembled using *Sequencher* software as explained in detail previously [7].

3. Results and Discussion

Figure 1 shows results related to averages of 100 × Δ_j values ($\bar{\Delta}$) derived from Equations (7) (*P*-I, black data) or (8) (*P*-II, red data) as a function of *n* (Figure 1(A)) and μ (Figure 1(B)). The least squares curve-fitting results show that the Figure 1(A) data follow the general form $\bar{\Delta}=\Delta_n\cdot n^{\bar{a}}$ whereupon $\bar{a}$ (averaged across 5 *n*-based fits) = −0.556 ± 0.00986 (black data sets; ±*s*) or $\bar{a}=-0.529\pm0.0387$ (red data). These findings suggest that $\bar{\Delta}$ changes as the inverse square root of *n* for all values of μ. Figure 1(C) displays these same results on a linearized scale (*X*-axis = $\sqrt[-2]{n}$) whereupon the slopes $\left(\partial\bar{\Delta}/\partial\left[\sqrt[-2]{n}\right]\right)\sim\Delta_n$. Figure 1(B) illustrates that the Δ_n values derived from Figure 1(A) non-linear regression change as the inverse square root of μ: *i.e.*, $\Delta_n=A\cdot\mu^a$ where *a* = −0.547 ± 0.0179 (black data) or −0.503 ± 0.0374 (red data); *a* ± *ASE*. Figure 1(D) shows Figure 1(B) results plotted on an appropriately linearized scale (*X*-axis = $\sqrt[-2]{\mu}$) as indicated by the above analysis whereupon the slope $\left(\partial\Delta_n/\partial\left[\sqrt[-2]{\mu}\right]\right)\sim A$. Combining results from Figure 1(A) and Figure 1(B) we see that $\bar{\Delta}\sim A\cdot\sqrt[-2]{n\cdot\mu}$. The average value for *A* was 0.804 ± 0.0460 (*P*-I & *P*-II curve-fitting results ±*s*).

Figure 2 displays *MPN*-based enumeration data, Equation (9), manipulated in a similar fashion as that of the above Poisson-based results with a nearly identical result. The least squares curve-fitting shows that the data in Figure 2(A) once again follow the general form $\bar{\Delta}=\Delta_n\cdot n^{\bar{a}}$ with $\bar{a}$ = −0.554 ± 0.0499 (±*s*) which is the average *a* from 5× μ-based data sets. Figure 2(C) shows these same findings graphed on a linearized scale ($X=\sqrt[-2]{n}$) whereupon the slopes = Δ_n. Figure 2(B) also shows that the Δ_n values, derived from Figure 2(A) non-linear regression, change as the inverse square root of μ: $\Delta_n=A\cdot\mu^a$ where *a* = −0.515 ± 0.0910 (±*ASE*). As previously observed, when these results are presented on a linearized scale ($X=\sqrt[-2]{\mu}$; Figure 2(D)) the slope is equivalent to the parameter *A*. Combining fitting results from Figure 2(A) and Figure 2(B) we again note that $\bar{\Delta}\sim A\cdot\sqrt[-2]{n\cdot\mu}$ (*A* = 0.807 ± 0.139; ± *ASE*).

Completely homologous relationships to the Poisson and *MPN* findings were also noted with Gaussian-based data (Figure 3) whereupon the least squares curve-fitting in Figure 3(A) shows that these data obey, again, the general form $\bar{\Delta}=\Delta_n\cdot n^{\bar{a}}$ whereupon $\bar{a}$ = −0.561 ± 0.0276 (±*s*; averaged across all σ since μ was fixed). Figure 3(C) has these same findings plotted on a linear scale ($X=\sqrt[-2]{\mu}$) where the slopes = Δ_n. Figure 3(B) and Figure 3(D) also show that the Δ_n values derived from Figure 3(A) and Figure 3(C) non-linear regression change linearly with $\sigma\div\mu$: *i.e.*, $\Delta_n=A\cdot\sigma\div\mu$ (*A* = 0.725 ± 0.0977; ± *ASE*). All Gaussian-based data fitting results combined indicate that

$\bar{\Delta} = A\cdot\sigma\cdot\sqrt[-2]{n}\cdot\mu^{-1} = A\cdot\sigma_{\bar{x}}\cdot\mu^{-1} = A\cdot C_V\left[\bar{x}\right]$ whereupon $C_V\left[\bar{x}\right]$ is the coefficient of variation for a population of means associated with *x*.

3.1. Equivalence of Sampling Errors Associated with Any PDF

The counting results alluded to above (*P-I*, *P-II*, & *MPN*) are similar to those observed previously: [5] [15] [16] *i.e.*, stochastic sampling errors associated with microbiological colony counting and *MPN* data are proportional to the inverse square root of $n\times\mu$. Also, the Poisson population-based results compare favorably with those obtained from actual colony counting experiments [14]. Thus, for all Poisson-based data (**Figure 1**)

$$\bar{\Delta} \propto \frac{1}{\sqrt{n\cdot\mu}} = \frac{1}{\sqrt{n}}\cdot\frac{\sqrt{\mu}}{\mu} = \frac{\sigma}{\sqrt{n}}\cdot\frac{1}{\mu} = \frac{\sigma_{\bar{x}}}{\mu} = C_V\left[\bar{x}\right] \tag{11}$$

because $\sigma = \sqrt{\mu}$. We have simplified the expression by utilizing the term $\sigma_{\bar{x}}$ [22] ($= \sigma \div \sqrt{n}$) which can be derived using the propagation of errors method [20]. Such nomenclature exemplifies the utilization of P_G, as an approximation for P_P, associated with a population of sample means ($\bar{x}$) of mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$. However, for *MPN* results, does $\sigma \sim \sqrt{\mu}$ as an approximation? This question is addressed in detail (**Figures 4-6**).

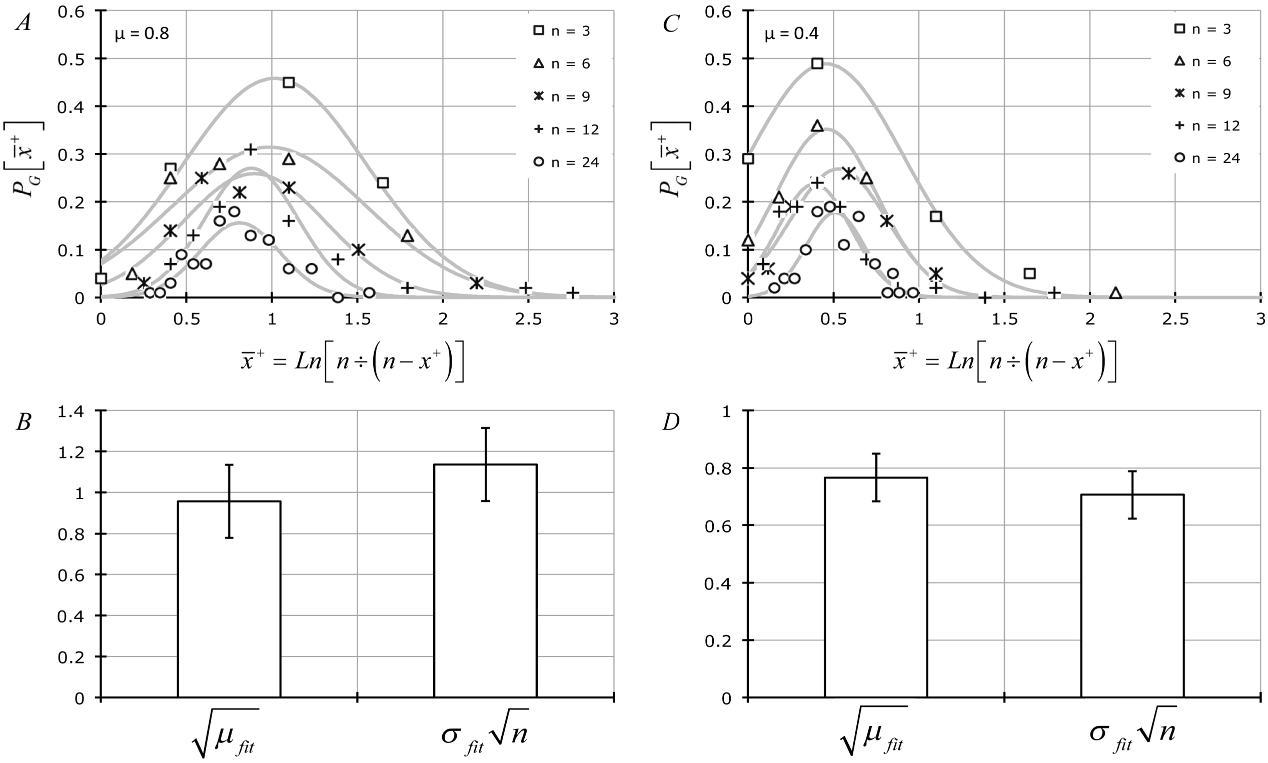


Figure 4. (A) & (C) Frequency of observing each set of *MPN*-based calculated number of entities per sample tested ($\bar{x}^+ = \ln\left[n\div\left(n-x^+\right)\right]$; μ = 0.8 for (A) & (B); μ = 0.4 for (C) & (D) fit to Equation (5) (*i.e.*, $P_G\left[\bar{x}^+\right]$ as a function of $\bar{x}^+$). (B) and (D) shows that $\sigma_{fit} \sim \sqrt{\mu_{fit}} \div \sqrt{n} \sim \sigma_{\bar{x}}$ (*i.e.*, for *MPN*, $\sigma = \sqrt{\mu}$) for all modeled *n*-samplings. Error bars are = $t_{0.05}\times$ the experimental (overall) $s_{\bar{x}} = \sqrt{EMS\div n}$.

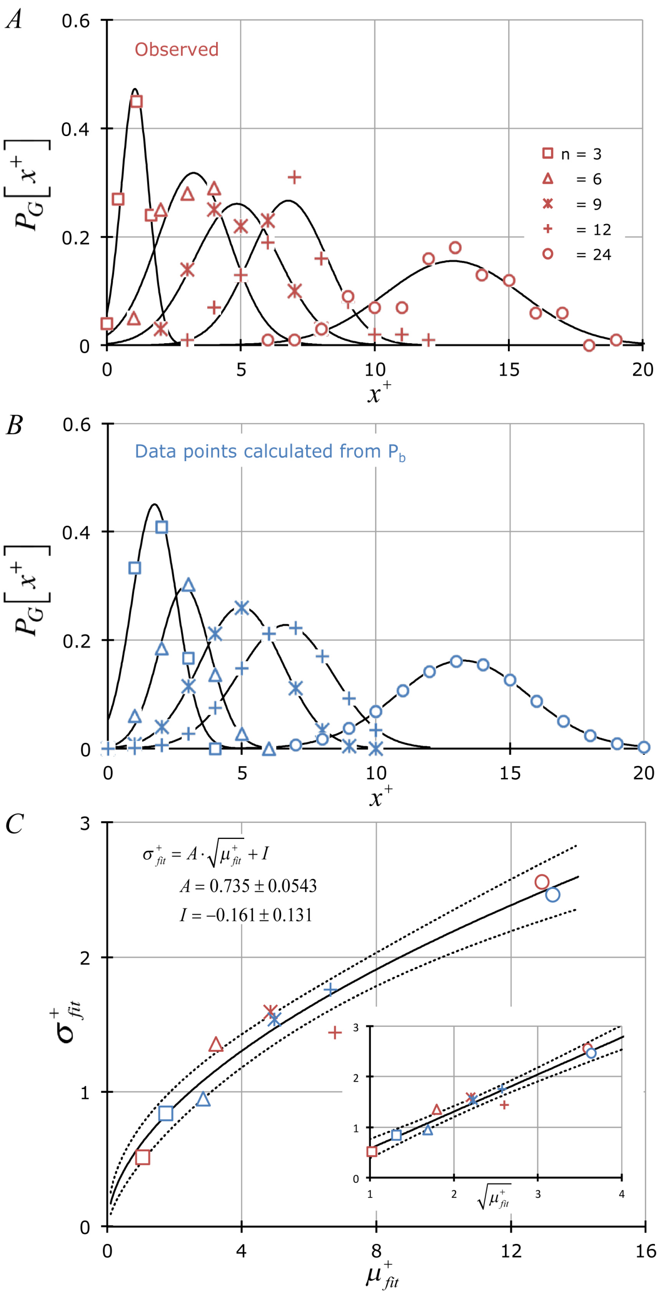


Figure 5. (A) & (B) Frequency of observing each set of *MPN*-based number of positive counts x^+ tested: $\mu = 0.8$ and $n = 3, 6, 9, 12, 24$; (A) data points [red] = frequency of observed x^+, (B) data points [blue] = calculated frequency of x^+ using *Mathematica*

$$P_b\left[x^+\right] = \text{Table}\left[N\left[\frac{\left(e^{-\mu}\right)^{n-x^+}\left(1-e^{-\mu}\right)^{x^+} n!}{\left(n-x^+\right)!x^+!}\right],\left\{x^+,0,n+1\right\}\right]$$

from P_b, Equation (3), fit to a Gaussian probability distribution: e.g., $P_G\left[x^+\right]$, Equation (5). (C) Demonstrates that $\sigma^+_{fit} \propto \sqrt{\mu^+_{fit}}$. Linear fit showing slope (A) and intercept (I) ± ASE. The non-linear fits were $\sigma^+_{fit} = A\cdot\left(\mu^+_{fit}\right)^{0.550\pm0.0463}$. Best fit curves shown ± $P = 0.05$ *CL*.

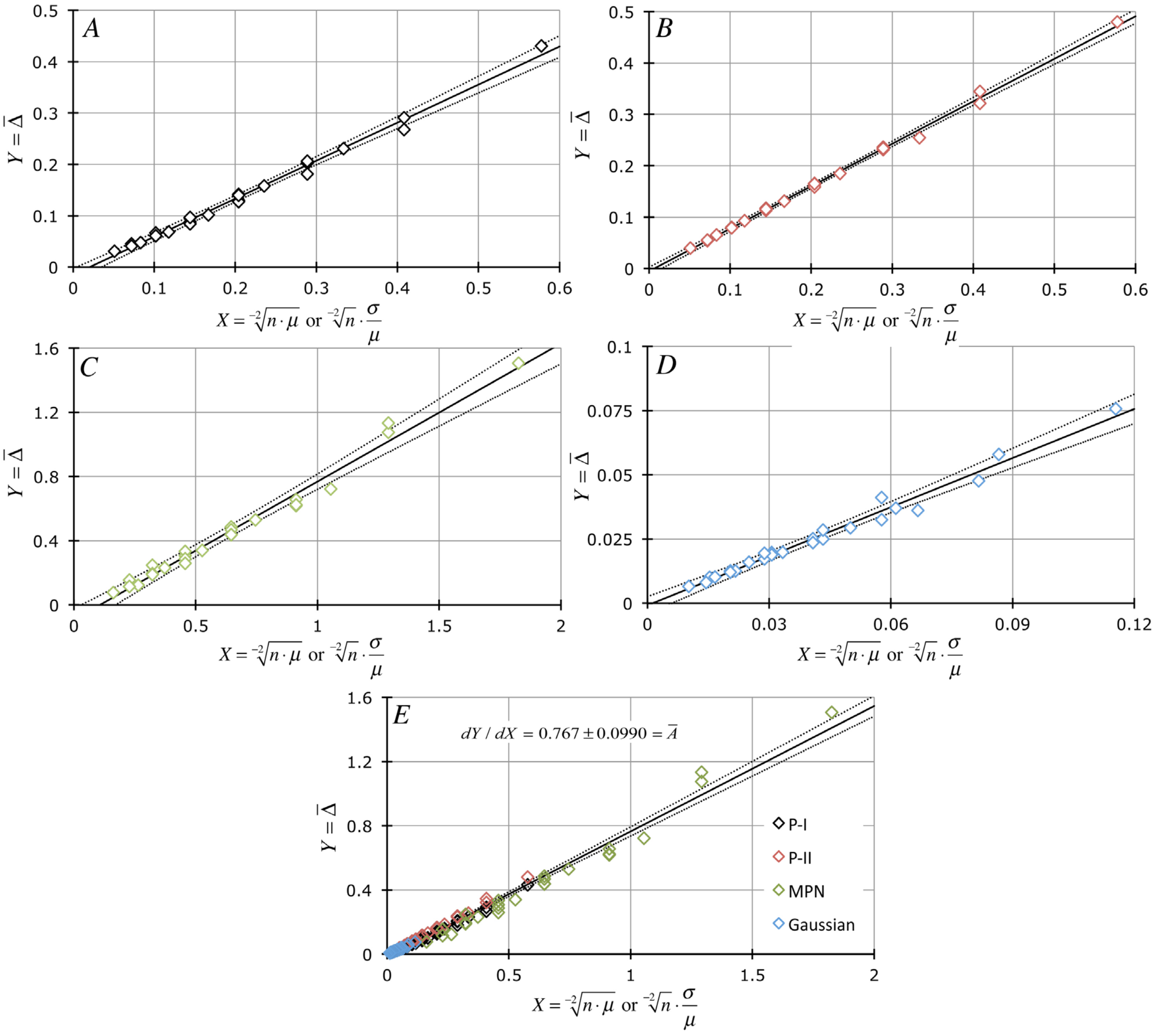


Figure 6. Demonstration that $\mathrm{d}\bar{\Delta}/\mathrm{d}\left(\sigma_{\bar{x}}/\mu\right) \sim A \sim \mathrm{d}\bar{\Delta}/\mathrm{d}\left(\sqrt[-2]{n\cdot\mu}\right)$. All data are plotted ± $P = 0.001$ *CL*. (A) is related to *P-I* data (A = 0.741 ± 0.0203; ± *ASE*). (B) is related to *P-II* data (A = 0.827 ± 0.0133). (C) is related to *MPN* data (A = 0.861 ± 0.0273). (D) is related to *Gaussian* data (A = 0.637 ± 0.0280). All data are merged in (E): slope of this relationship which involves all three *PDF*s is 0.767 ± 0.0990.

In Figure 4(A) and Figure 4(C), we have examined some of our *MPN* data (μ = 0.8 per sample in Figure 4(A) and μ = 0.4 per sample in Figure 4(C) at the various levels of n-sampling) by converting the total number of positive occurrences (x_j^+) in n observations to the most probable number of entities in the hypothetical sampled aliquot ($\bar{x}_j^+ = \ln\left[n/\left(n-x_j^+\right)\right]$) and curve-fit the frequency of occurrence of each $\bar{x}_j^+$ to Gaussian *PDF*s (Equation (5); $P_G\left[\bar{x}^+\right]$). From these curve fits we extracted the parameters σ_{fit} and μ_{fit}. In Figure 4(B) and Figure 4(D) we show that the average $\sigma_{fit}\sqrt{n} \sim \sqrt{\mu_{fit}}$ (*i.e.*, $\sigma_{fit} = \sqrt{\mu_{fit}} \div \sqrt{n} = \sigma_{\bar{x}}$) and, therefore, $\sigma = \sqrt{\mu}$. This finding indicates that Equation (11) can be applied to both Poisson and *MPN* results as a reasonable approximation. We have confirmed the *MPN* results in Figure 2 and Figure 4 by showing that the frequency distribution of x^+ which we have observed in these experiments closely follows Equation (3) (compare Figure 5(A) with Figure 5(B)) whereupon we establish that σ^+_{fit}, the standard deviation associated with the distribution of x^+ via the Gaussian approximation, was proportional to $\sqrt{\mu^+_{fit}}$ (Figure 5(C)) for both observed (red data) and calculated (blue data) x^+ with a proportionality constant numerically similar to A (=0.735 ± 0.0543; ±*ASE*) alluded to above.

The equality in Equation (11) is also visually confirmed by the results shown in Figure 6 where one can see that all values of $\bar{\Delta}$ closely follow the linear expression $\bar{\Delta} = A \cdot X$ (for $X = \sigma_{\bar{x}} \div \mu$ or $\sqrt[-2]{n \cdot \mu}$; A = 0.781 ± 0.0107; ±*ASE*) showing that

$$\frac{\partial\bar{\Delta}}{\partial\left[\sqrt[-2]{n\cdot\mu}\right]} = \frac{\partial\bar{\Delta}}{\partial\left[\sigma_{\bar{x}} \div \mu\right]}.$$

Since the combined data in Figure 6 are linear with a near-zero intercept (−0.0168 ± 0.00443), then

$$\frac{\bar{\Delta}}{\sqrt[-2]{n\cdot\mu}} = \frac{\bar{\Delta}}{\sigma_{\bar{x}} \div \mu}$$

therefore cross-multiplying gives

$$\bar{\Delta}\cdot\sigma_{\bar{x}} \div \mu = \bar{\Delta}\cdot\sqrt{n\cdot\mu}$$

and dividing both sides by $\bar{\Delta}$ produces the equality

$$\sigma_{\bar{x}} \div \mu = \sqrt[-2]{n\cdot\mu}.$$

All sampling error-related findings are summarized in Figure 7.

3.2. Demonstration That $A = \partial s_{\Delta_j} / \partial\bar{\Delta} = C_V\left[\Delta_j\right]$

Lastly, all these assertions are substantiated by the observation (Figure 8) that the standard deviations associated with *all* our sampling error measurements (s_{Δ_j}) change linearly as a function of the 4 (*P-I*, *P-II*, *MPN*, Gaussian) sets of $\bar{\Delta}$ data with an average slope (*i.e.*, average of the 4 $\partial s_{\Delta_j} / \partial\bar{\Delta}$ values = 0.716 ± 0.0739) equivalent to the various values for A in Figures 1-3, Figure 5 and Figure 6. In fact, the slope in Figure 8 defines the coefficient of variation in $\bar{\Delta}$ ($C_V\left[\Delta_j\right]$) and, if equal to A, then

$$\frac{\partial s_{\Delta_j}}{\partial\bar{\Delta}} = \frac{\partial\bar{\Delta}}{\partial X} \tag{12}$$

where X = either $\sqrt[-2]{n\cdot\mu}$ or $\sigma_{\bar{x}} \div \mu$. Since s_{Δ_j} in Figure 8 and $\bar{\Delta}$ in Figure 6 are linear functions with a near zero intercept then, assuming Equation (12) is true,

$$\frac{s_{\Delta_j}}{\bar{\Delta}} = \frac{\bar{\Delta}}{X}.$$

Substituting $\bar{\Delta}$ with $A\cdot X$

$\bar{\Delta}$			
P-I	*P-II*	*MPN*	Gaussian
$\bar{\Delta} = J^{-1} \cdot \sum_{j=1}^{J=100} \frac{\left\|\sqrt{\mu} - s_j\right\|}{\mu}$	$\bar{\Delta} = J^{-1} \cdot \sum_{j=1}^{J=100} \frac{\left\|\mu - \bar{x}_j\right\|}{\mu}$	$\bar{\Delta} = J^{-1} \cdot \sum_{j=1}^{J=100} \frac{\left\|\mu^+ - x_j^+\right\|}{\mu^+}$	$\bar{\Delta} = J^{-1} \cdot \sum_{j=1}^{J=100} \frac{\left\|\sigma - s_j\right\|}{\mu}$
$\bar{\Delta}_{fit} = \frac{\Delta_n}{\sqrt{n}}$	$\bar{\Delta}_{fit} = \frac{\Delta_n}{\sqrt{n}}$	$\bar{\Delta}_{fit} = \frac{\Delta_n}{\sqrt{n}}$	$\bar{\Delta}_{fit} = \frac{\Delta_n}{\sqrt{n}}$
$\Delta_{n,fit} = \frac{A}{\sqrt{\mu}}$	$\Delta_{n,fit} = \frac{A}{\sqrt{\mu}}$	$\Delta_{n,fit} = \frac{A}{\sqrt{\mu}}$	$\Delta_{n,fit} = A\frac{\sigma}{\mu}$
combined	combined	combined	combined
$\bar{\Delta}_{fit} = \frac{A}{\sqrt{n \cdot \mu}}$	$\bar{\Delta}_{fit} = \frac{A}{\sqrt{n \cdot \mu}}$	$\bar{\Delta}_{fit} = \frac{A}{\sqrt{n \cdot \mu}}$	$\bar{\Delta}_{fit} = A \cdot \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{\mu}$
$= A \cdot \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{\mu}}{\mu}$	$= A \cdot \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{\mu}}{\mu}$	$= A \cdot \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{\mu}}{\mu}$ *	$= A \cdot \frac{\sigma_{\bar{x}}}{\mu}$
$= A \cdot \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{\mu}$	$= A \cdot \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{\mu}$	$= A \cdot \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{\mu}$	
$= A \cdot \frac{\sigma_{\bar{x}}}{\mu}$	$= A \cdot \frac{\sigma_{\bar{x}}}{\mu}$	$= A \cdot \frac{\sigma_{\bar{x}}}{\mu}$	
$A = 0.771 \pm 0.0123$	$A = 0.836 \pm 0.0296$	$A = 0.801 \pm 0.0910$	$A = 0.725 \pm 0.0977$

* because

Fig. 4: $P[\bar{x}]$ dependence upon $\bar{x}$ ($\bar{x} = Ln[n \div (n - x^+)]$) $: \sigma_{fit} \sim \sqrt{\mu_{fit}} \div \sqrt{n}$

Fig. 5: $P[x^+]$ dependence upon x^+ : $\sigma^+ \propto \sqrt{\mu^+}$

Figure 7. Summary of curve-fitting results associated with each *PDF* and method for calculating empirical stochastic sampling error ($\bar{\Delta}$). Each constant of proportionality A is presented ± *ASE*. For binomial data (*MPN*) $\mu = V \cdot \delta$ (the population average number of entities in V) and $n \cdot P^+ = \mu^+$ (the population average number of positive responses out of n observations).

$$\frac{s_{\Delta_j}}{A \cdot X} = \frac{A \cdot X}{X}$$

$$\frac{s_{\Delta_j}}{X} = A^2$$

$$s_{\Delta_j} = A^2 \cdot X = A(A \cdot X) = A \cdot \bar{\Delta}$$

and therefore

$$\frac{s_{\Delta_j}}{\bar{\Delta}} = C_V\left[\Delta_j\right] = A$$

The above equality establishes that the coefficient of variation associated with $\bar{\Delta}$ ($C_V\left[\Delta_j\right]$) is equivalent to the proportionality constant A seen in Figures 1-3 and Figure 6. Thus sampling errors can be estimated from the relationship $\bar{\Delta} = C_V\left[\Delta_j\right] \times C_V\left[\bar{x}\right]$ whereupon $C_V\left[\Delta_j\right] \sim 0.75$ for all *PDF*s we have tested.

3.3. Minimized Errors Associated with a Well-Sampled Food Microbiome via Most Probable Composition [7]

Based upon these results, the estimation of $C_V\left[\bar{x}\right]$ (*i.e.*, $s_{\bar{x}} \div \bar{x}$) should be germane in determining if data have been appropriately sampled. Figure 9 illustrates that all stochastic errors associated with native aerobic bacteria surviving

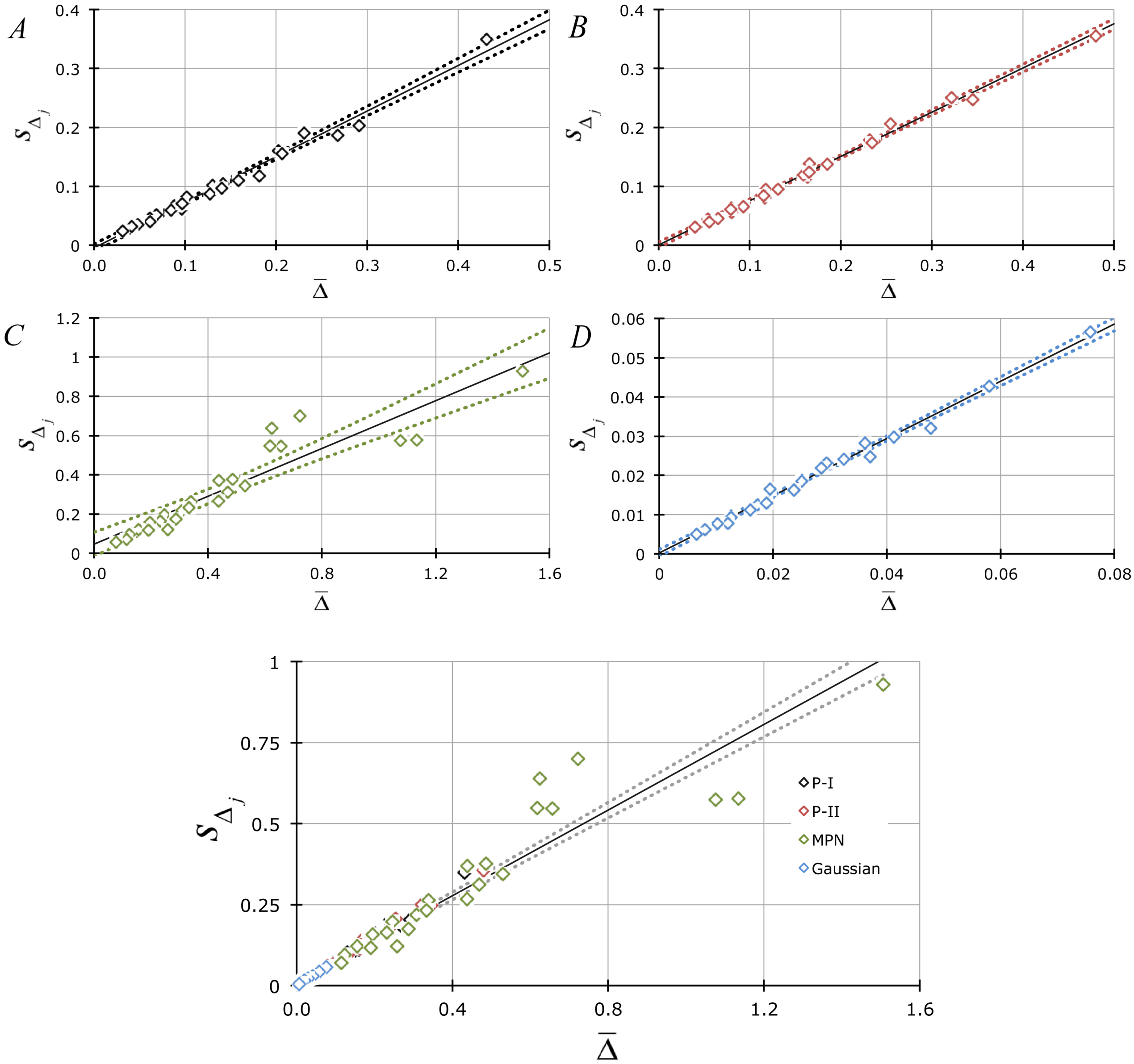
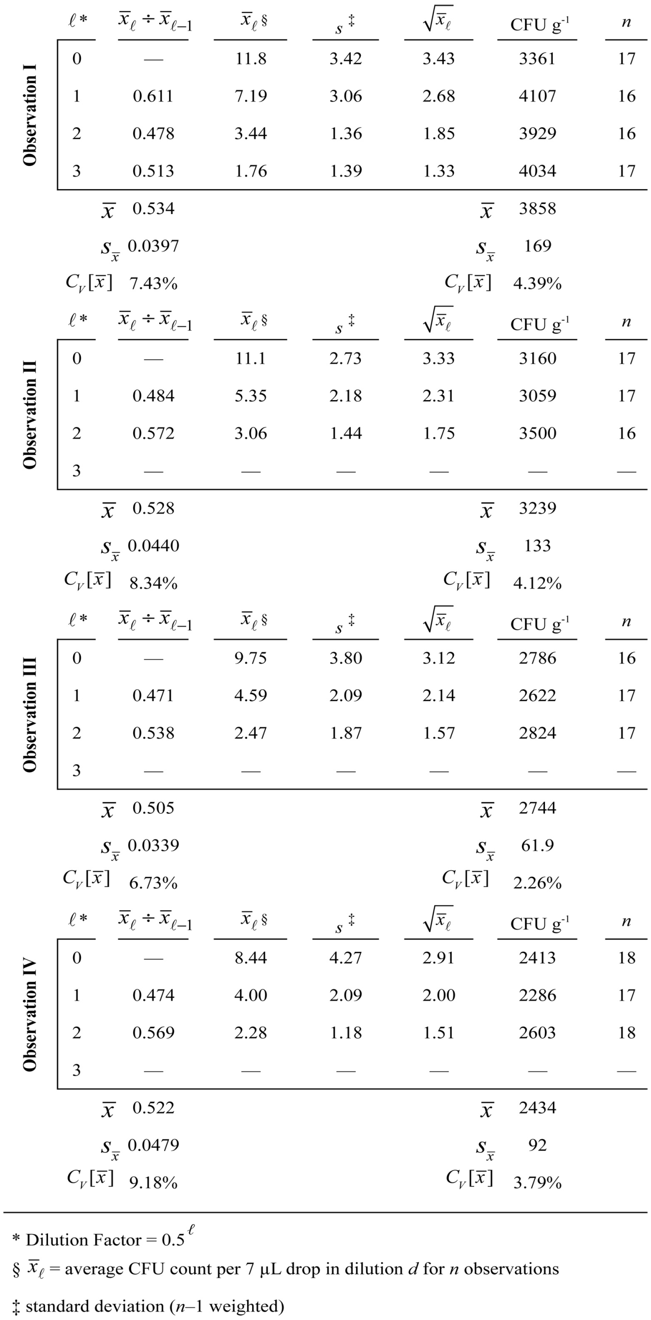


Figure 8. (A)-(D): Dependency of the standard deviation (plotted ± P = 0.05 confidence limits) derived from each experimental Δ_j array ($Y = s_{\Delta_j}$; $j = 1,2,\cdots,25$) on their averages ($X = \bar{\Delta}$): Figure 8(A) = *P-I* data (Spearman's coefficient of rank correlation: [22] $\rho_S = 0.996$; $P \ll 10^{-3}$); Figure 8(B) = *P-II* data ($\rho_S = 0.988$; $P \ll 10^{-3}$); Figure 8(C) = *MPN* data ($\rho_S = 0.979$; $P \ll 10^{-3}$); Figure 8(D) = Gaussian data ($\rho_S = 0.994$; $P \ll 10^{-3}$). The average slopes associated with these 4 relationship = 0.716 ± 0.0739 (± *s*). All points (25× $\bar{\Delta}$ per set) from (A) through (D) are combined in the bottom-most figure ($\mathrm{d}s_{\Delta_j}/\mathrm{d}\bar{\Delta} = 0.661 \pm 0.0186$; ± *ASE*). The value $\mathrm{d}s_{\Delta_j}/\mathrm{d}\bar{\Delta}$ is equivalent to an experimental coefficient of variation for $\bar{\Delta} = C_V\left[\Delta_j\right]$.

on commercially available, frozen vegetables were sufficiently sampled using an n = 16 - 18 inasmuch as the $C_V\left[\bar{x}\right]$-values associated with the normalized colony counts ($CFU\,\mathrm{g}^{-1}$ averaged across all ℓ dilutions = $\bar{x}_\ell$ ÷ 0.007 mL per drop ÷ 0.5^ℓ dilution factor × 57.2 mL total original sample volume ÷ 28.6 g total frozen vegetable mass) were appropriately small (ranging between *ca.* 2% to 4%). In a



	ℓ*	$\bar{x}_\ell \div \bar{x}_{\ell-1}$	$\bar{x}_\ell$ §	s ‡	$\sqrt{\bar{x}_\ell}$	CFU g^{-1}	n
Observation I	0	—	11.8	3.42	3.43	3361	17
	1	0.611	7.19	3.06	2.68	4107	16
	2	0.478	3.44	1.36	1.85	3929	16
	3	0.513	1.76	1.39	1.33	4034	17
	$\bar{x}$	0.534			$\bar{x}$	3858	
	$s_{\bar{x}}$	0.0397			$s_{\bar{x}}$	169	
	$C_V[\bar{x}]$	7.43%			$C_V[\bar{x}]$	4.39%	

	ℓ*	$\bar{x}_\ell \div \bar{x}_{\ell-1}$	$\bar{x}_\ell$ §	s ‡	$\sqrt{\bar{x}_\ell}$	CFU g^{-1}	n
Observation II	0	—	11.1	2.73	3.33	3160	17
	1	0.484	5.35	2.18	2.31	3059	17
	2	0.572	3.06	1.44	1.75	3500	16
	3	—	—	—	—	—	—
	$\bar{x}$	0.528			$\bar{x}$	3239	
	$s_{\bar{x}}$	0.0440			$s_{\bar{x}}$	133	
	$C_V[\bar{x}]$	8.34%			$C_V[\bar{x}]$	4.12%	

	ℓ*	$\bar{x}_\ell \div \bar{x}_{\ell-1}$	$\bar{x}_\ell$ §	s ‡	$\sqrt{\bar{x}_\ell}$	CFU g^{-1}	n
Observation III	0	—	9.75	3.80	3.12	2786	16
	1	0.471	4.59	2.09	2.14	2622	17
	2	0.538	2.47	1.87	1.57	2824	17
	3	—	—	—	—	—	—
	$\bar{x}$	0.505			$\bar{x}$	2744	
	$s_{\bar{x}}$	0.0339			$s_{\bar{x}}$	61.9	
	$C_V[\bar{x}]$	6.73%			$C_V[\bar{x}]$	2.26%	

	ℓ*	$\bar{x}_\ell \div \bar{x}_{\ell-1}$	$\bar{x}_\ell$ §	s ‡	$\sqrt{\bar{x}_\ell}$	CFU g^{-1}	n
Observation IV	0	—	8.44	4.27	2.91	2413	18
	1	0.474	4.00	2.09	2.00	2286	17
	2	0.569	2.28	1.18	1.51	2603	18
	3	—	—	—	—	—	—
	$\bar{x}$	0.522			$\bar{x}$	2434	
	$s_{\bar{x}}$	0.0479			$s_{\bar{x}}$	92	
	$C_V[\bar{x}]$	9.18%			$C_V[\bar{x}]$	3.79%	

* Dilution Factor = 0.5^{ℓ}

§ $\bar{x}_\ell$ = average CFU count per 7 μL drop in dilution d for n observations

‡ standard deviation (n–1 weighted)

Figure 9. Estimation of the stochastic sampling errors ($\bar{x}_\ell \div \bar{x}_{\ell-1} \sim$ calculated dilution factors; $s \sim \sqrt{\bar{x}_\ell}$; $C_V[\bar{x}]$ for all counts~4% across all dilutions ℓ) associated with a well-sampled [15] (n = 16 - 18) Poisson population (native bacteria on frozen vegetables: 28.6 grams rinsed with 57.2 mL *PBS* + Tween 20). All the colonies in ℓ = 2 (Control & grown at 30˚C = 55 colonies; Hollow Fiber Concentrated & grown at 30˚C = 49 colonies; Control & grown at 37˚C = 41 colonies; Hollow Fiber Concentrated & grown at 37˚C = 41 colonies) were collected and identified using 16S *rDNA* Sanger sequencing (*EubA* and *EubB* primers) as described previously [7]. Bacterial compositions were nearly identical for all samplings and treatment combinations.

similar vein, it is pertinent that the observed (*s*) and calculated ($\sqrt{\overline{x}_\ell}$) standard deviations associated with the counts per drop were equivalent since the average deviation ($\left|s-\sqrt{\overline{x}_\ell}\right|$) from ideality varied only 15.7% ± 3.54% ($\pm s_{\overline{x}}$). Lastly it is also significant that the dilution factors calculated from the ratios of average plate counts ($\overline{x}_\ell \div \overline{x}_{\ell-1}$) were very close to ½ (average 0.523 ± 0.0172) which also argues for a minimized Δ.

Across the 4 observational sets (I, II, III, and IV) depicted in **Figure 9**, the total number of collected colonies (from $\ell = 2$) was 55 (*n* = 16), 49 (*n* = 16), 42 (*n* = 17), and 41 (*n* = 18), respectively. Bacteria identifications for each of these colonies were based upon *rDNA* sequence matching 1200 - 1400 basepair contigs searching against *NCBI*'s GenBank database. The *rRNA* "gene" sequencing results for the 2 major isolates (making up 88.3% ± 3.28% of the total sampled colonies) show that the 4 sets of observed bacterial compositions were nearly identical (43.6% ± 8.05% *Luconostoc* and 44.6% ± 13.3% *Lactococcus*; ±*s*) [23]. The remainder of the colonies was mainly *Acinetobacter* (3.74% ± 3.34%) and *Streptococcus* (4.17% ± 2.75%) with small amounts of diverse isolates (e.g., *Staphylococcus*, *Arthrobacter*, *Sphingobacterium*, *Enterococcus*, *Kocuria*, *Raoultella*, and *Bacillus*: averaging 1.49% ± 1.09% each). Such variability is expected for the relatively rare isolates (≤4%) due to errors associated with random sampling. The two *major* species sampled were relatively repeatable because of their abundance, adequate sampling, and very little treatment effect. The *minor* constituents would have to have been sampled 2.77 ± 0.647-fold more (*n* > 44) for an equivalent accuracy to the *Luconostoc* and *Lactococcus* fractions since the requisite number of samplings for the low count fractions, above, is proportional to the inverse cube root [5] [16] of the number of counts per sampled volume (~ $\sqrt[3]{\overline{x}_{major}} \div \sqrt[3]{\overline{x}_{minor}}$).

4. Summary

We have performed analyses associated with empirical stochastic sampling errors linked to data generated from 3 common probability density functions. We have used these to describe the limiting behavior of Δ by generating models which suggest a generalized, and facile, mathematical solution. Based upon all our experiments, the common algebraic solution, regardless of parent distribution, is that experimental sampling errors are proportional to $\sigma_{\overline{x}} \div \mu$. This generalized relationship is intuitively reasonable inasmuch as this is the C_V *for any population of sample means* ($C_V\left[\overline{x}\right]$) and describes how closely $\overline{x}$ values approach *μ* as *n* increases. The proportionality constant for all these findings was found to be mathematically related to $C_V[\Delta_j]$ or $\partial s_{\Delta_j} \big/ \partial\overline{\Delta}$, which is the coefficient of variation associated with the error measurement itself. Lastly, using estimates of these sampling-associated errors ($C_V\left[\overline{x}\right] \sim s_{\overline{x}} \div \overline{x}$), we show that when a test microbiome was sufficiently sampled, several measures of stochastic sampling error were reasonably small for both counting and *DNA* sequence-based results.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Halvorson, H.O. and Ziegler, N.R. (1933) Application of Statistics to Problems in Bacteriology. I. A Means of Determining Bacterial Population by the Dilution Method. *Journal of Bacteriology*, **25**, 101-121.

[2] Kubitschek, H.E. (1990) Cell Volume Increase in *Escherichia coli* after Shifts to Richer Media. *Journal of Bacteriology*, **172**, 94-101. https://doi.org/10.1128/jb.172.1.94-101.1990

[3] Barkworth, H. and Irwin, J.O. (1938) Distribution of Coliform Organisms in Milk and the Accuracy of the Presumptive Coliform Test. *Journal of Hygiene*, **38**, 446-457. https://doi.org/10.1017/S0022172400011311

[4] Best, D.J. (1990) Optimal Determination of Most Probable Numbers. *International Journal of Food Microbiology*, **11**, 159-166. https://doi.org/10.1016/0168-1605(90)90051-6

[5] Irwin, P., Reed, S., Nguyen, L., Brewster, J. and He, Y. (2013) Non-Stochastic Sampling Error in Quantal Analyses for *Campylobacter* Species on Poultry Products. *Analytical and Bioanalytical Chemistry*, **405**, 2353-2369. https://doi.org/10.1007/s00216-012-6659-2

[6] Irwin, P., Gehring, A., Tu, S.-I., Brewster, J., Fanelli, J. and Ehrenfeld, E. (2000) Minimum Detectable Level of Salmonellae Using a Binomial-Based Ice Nucleation Detection Assay. *Journal of AOAC International*, **83**, 1087-1095.

[7] Irwin, P.L., Nguyen, L.-H.T., Chen, C.-Y. and Paoli, G. (2008) Binding of Nontarget Microorganisms from Food Washes to Anti-*Salmonella* and anti-*E. coli* O157 Immunomagnetic Beads: Most Probable Composition of Background Eubacteria. *Analytical and Bioanalytical Chemistry*, **391**, 525-536. https://doi.org/10.1007/s00216-008-1959-2

[8] de St. Groth, S.F. (1982) The Evaluation of Limiting Dilution Assays. *Journal of Immunological Methods*, **49**, R11-R23. https://doi.org/10.1016/0022-1759(82)90269-1

[9] Bevington, P.R. and Robinson, D.K. (1992) Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, Boston, 17-23 and 41-43.

[10] Irwin, P., Fortis, L. and Tu, S.-I. (2001) A Simple Maximum Probability Resolution Algorithm for Most Probable Number Analysis Using Microsoft Excel. *Journal of Rapid Methods and Automation in Microbiology*, **9**, 33-51. https://doi.org/10.1111/j.1745-4581.2001.tb00226.x

[11] Gosset, W.S. (1907) "Student" on the Error of Counting with a Haemocytometer. *Biometrika*, **5**, 351-360. https://doi.org/10.1093/biomet/5.3.351

[12] Fisher, R.A. (1922) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society*, *London*, *Series A*, **222**, 309-368. https://doi.org/10.1098/rsta.1922.0009

[13] Irwin, P.L., Nguyen, L.-H.T., Paoli, G.C. and Chen, C.-Y. (2010) Evidence for a Bimodal Distribution of *Escherichia coli* Doubling Times below a Threshold Initial Cell Concentration. *BMC Microbiology*, **10**, 207.

[14] Chen, C.-Y., Nace, G.W. and Irwin, P.L. (2003) A 6×6 Drop Plate Method for Si-

multaneous Colony Counting and MPN Enumeration of *Campylobacter jejuni*, *Listeria monocytogenes*, and *Escherichia coli*. *Journal of Microbiological Methods*, **55**, 475-479. https://doi.org/10.1016/S0167-7012(03)00194-5

[15] Irwin, P.L., Nguyen, L.-H.T. and Chen, C.-Y. (2008) Binding of Nontarget Microorganisms from Food Washes to Anti-*Salmonella* and Anti-*E. coli* O157 Immunomagnetic Beads: Minimizing the Errors of Random Sampling in Extreme Dilute Systems. *Analytical and Bioanalytical Chemistry*, **391**, 515-524. https://doi.org/10.1007/s00216-008-1961-8

[16] Irwin, P.L., Nguyen, L.-H.T. and Chen, C.-Y. (2010) The Relationship between Purely Stochastic Sampling Error and the Number of Technical Replicates Used to Estimate Concentration at an Extreme Dilution. *Analytical and Bioanalytical Chemistry*, **398**, 895-903. https://doi.org/10.1007/s00216-010-3967-2

[17] Trotter, H.F. (1959) An Elementary Proof of the Central Limit Theorem. *Archiv der Mathematik*, **10**, 226-234. https://doi.org/10.1007/BF01240790

[18] Hartley, H.O. (1961) The Modified Gauss-Newton Method for Fitting of Non-Linear Regression Functions by Least Squares. *Technometrics*, **3**, 269-280. https://doi.org/10.1080/00401706.1961.10489945

[19] Irwin, P.L., Damert, W.C. and Doner, L.W. (1994) Curve Fitting in Nuclear Magnetic Resonance Spectroscopy: Illustrative Examples Using a Spreadsheet and Microcomputer. *Concepts in Magnetic Resonance*, **6**, 57-67. https://doi.org/10.1002/cmr.1820060105

[20] Beers, Y. (1957) Introduction to the Theory of Error. Addison-Wesley Publishing Company, Inc., Reading, 29-30.

[21] Salter, C. (2000) Error Analysis Using the Variance-Covariance Matrix. *Journal of Chemical Education*, **77**, 1239-1243. https://doi.org/10.1021/ed077p1239

[22] Steel, R.G.D. and Torrie, J.H.D. (1960) Principles and Procedures of Statistics. McGraw-Hill, New York, 409.

[23] Irwin, P., Capobianco, J., Nguyen, L., He, Y., Gehring, M., Gehring, A. and Chen, C.-Y. (2019) Bacterial Cell Recovery after Hollow Fiber Microfiltration Sample Concentration and Washing: Most Probable Bacterial Composition in Frozen Vegetables.

Definitions

Indices = $i\ (=1,2,\cdots,n)$ observations per experiment; $j\ (=1,2,\cdots,J=100)$ experiments with n observations each; $k\ (=1,2,\cdots,K)$ rows of X-Y values; ℓ $(=1,2,\cdots,L)$ dilutions; $m\ (=1,2,\cdots,M)$ iterations; $p\ (=1,2,\cdots,P)$ parameters

Δ_j = j^{th} experimental measure of sampling error out of J = 100 experiments: Equations (7)-(10).

$\overline{\Delta}$ = average sampling error in J = 100 observations of Δ_j

A = proportionality constant associated with $\overline{\Delta}$ curve-fitting to n, μ (or σ)

s_{Δ_j} = standard deviation associated with Δ_j measurement; for this work there are 25 ($n\times\mu$ or $n\times\sigma$ for the Gaussian populations) such s_{Δ_j} for each *PDF* type (2 types of Poisson, *MPN* or binomial, Gaussian)

μ = for either Poisson *PDF* or *MPN* assays ($\mu = V\cdot\delta$), the population average number of biological entities, or other analytes, per test; for Gaussian *PDF*, the population's average of any real-valued, randomly changing variable

V = the sample volume to be tested

V_e = volume of the biological entity, or other analyte, being tested

δ = concentration of the biological entity (count ÷ V) or other analyte

μ^+ = population average number of positive growth responses (*MPN*) out of n observations; $\mu^+ = n\cdot P^+$

σ^+ = the standard deviation associated with the probability density of x^+; the Gaussian approximations for σ^+ are plotted in **Figure 5(C)** as a function of Gaussian best fits for μ^+

P^- = probability that V_e will NOT contain the biological entity, or other analyte, being tested

P^+ = probability that V_e will contain the biological entity, or other analyte, being tested; $P^+ = 1-P^-$; Equation (1)

$$\partial_X f[X] = \partial f[X]/\partial X$$

x_{ij} = for Poisson populations, the i^{th} observation's number of counts per tested volume, surface area, etc. for each j^{th} experiment; for Gaussian populations, any real-valued, randomly changing variable

$$\overline{x}_j = \frac{1}{n}\cdot\sum\nolimits_{i=1}^{n} x_{ij}$$

x_j^+ = j^{th} experiment's number of positive growth responses out of n observations; $x_j^+ = \sum\nolimits_{i=1}^{n}\theta_{ij}$ where θ = 1 (positive) or 0 (negative)

$\overline{x}_j^+$ = j^{th} experiment's number of positive counts in V volume; $\overline{x}_j^+ = \ln\left[n\div\left(n-x_j^+\right)\right]$; the x-bar symbol is used here because this relations contains a parameter, x_j^+, which is the result of a summation across all θ_{ij}; it just isn't normalized to n

n = number of technical replicates in each j^{th} experiment; for *MPN*, number of observations each of volume V; for Poisson populations we have found [15] that the minimal number of replicates per assay was $n_{calc} = n_{\mu\to1}\cdot\sqrt[-3]{\mu}$ where $n_{\mu\to1}$ is the number of replicates necessary to enumerate a population with μ = 1

σ = population standard deviation associated with μ

$\sigma_{\bar{x}}$ = standard deviation of a population of sample means ($\bar{x}$); the formula for the $\sigma_{\bar{x}}$ statistic can be derived from the propagation of errors method [20] without covariance

$$\sigma_{\bar{x}} = \sqrt{\left(\frac{\partial \bar{x}}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial \bar{x}}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \cdots + \left(\frac{\partial \bar{x}}{\partial x_n}\right)^2 \sigma_{x_n}^2}$$
$$= \sqrt{n\frac{\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$$

since

$$\frac{\partial \bar{x}}{\partial x_1} = \frac{\partial \bar{x}}{\partial x_2} = \cdots = \frac{\partial \bar{x}}{\partial x_n} = \frac{1}{n}$$

and

$$\sigma_{x_1}^2 = \sigma_{x_2}^2 = \cdots = \sigma_{x_n}^2 = \sigma^2 .$$

s_j = any j^{th} experiment's estimation of population standard deviation

$s_{\bar{x}}$ = estimation of $\sigma_{\bar{x}}$ from a limited number of $\bar{x}_j$; $s_{\bar{x}} = s_j \div \sqrt{n}$

$C_V\left[\bar{x}\right]$ = coefficient of variation for a population of means; $C_V\left[\bar{x}\right] = \sigma_{\bar{x}} \div \mu_{\bar{x}} = \sigma_{\bar{x}} \div \mu$ estimated as $s_{\bar{x}} \div \bar{x}$

$C_V\left[x\right]$ = coefficient of variation for any set of observations *x*; $C_V\left[x\right] = \frac{\sigma}{\mu}$ estimated as $\frac{s}{\bar{x}}$

$C_V\left[\Delta_j\right]$ = $\partial s_{\Delta_j} \big/ \partial \bar{\Delta} \sim s_{\Delta_j} \div \bar{\Delta}$ if the s_{Δ_j} vs. $\bar{\Delta}$ intercept ~ 0

CLT = central limit theorem: the mean ($\mu_{\bar{x}}$) of a population of observed means ($\bar{x}$) will be approximately equal to the mean of the sampled population (μ) and the standard deviation of this population of means will be approximately equal to $\sigma_{\bar{x}}$; Equation (5) with $x = \bar{x}$, $\mu = \mu_{\bar{x}} = \mu$, and $\sigma = \sigma_{\bar{x}}$

PDF = probability density function or probability distribution function

P_b = binomial *PDF*: Equation (3)

P_P = Poisson *PDF*: Equation (4)

P_G = Gaussian *PDF*: Equation (5)

CL = confidence limit = *t*-statistic × s_{f_k} = $t \cdot s_{f_k}$

ASE = asymptotic standard error [19]; for any fitting parameter ω, $ASE = s_\omega = \sqrt{s_Y^2 \cdot \left[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right]_{\omega\omega}^{-1}}$; s_Y^2 = residual sum of squares ÷ (*K* – *M*) where *M* = the number of fitting parameters π_p ($p = 1, 2, \cdots, P$)

s_{f_k} = k^{th} row standard error of fitting function f_k; $s_{f_k} = \sqrt{s_Y^2\left(\boldsymbol{Z}_k\left[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right]^{-1}\boldsymbol{Z}_k^{\mathrm{T}}\right)}$

$\boldsymbol{Z}$ = partial first derivative matrix of f_k with respect to associated fitting parameters $\pi_1, \pi_2, \cdots, \pi_P$

$\boldsymbol{Z}^{\mathrm{T}}$ = transposition of $\boldsymbol{Z}$

$\boldsymbol{Z}_k$ = $\left[\partial_{\pi_1} f_k \quad \partial_{\pi_2} f_k\right]$ for $f_k = f\left[X_k; \pi_p\right]$