AMERICAN UNIVERSITY OF BEIRUT

A sufficient normality condition for Turing's formula

by

Fréderic Michael El Bayeh

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Mathematics
to the Department of Mathematics
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
April 2017

AMERICAN UNIVERSITY OF BEIRUT

A sufficient normality condition for Turing's formula

by

Fréderic Michael El Bayeh

Approved by:

____

Dr. Abbas ALHakim, Associate Professor — Advisor

Mathematics

Abbas Alhakim

Dr. Nabil Nassif, Professor — Member of Committee

Mathematics

Nabil Nassif

Dr. Stefano Monni, Assistant Professor — Member of Committee

Mathematics

Stefano Monni

Date of thesis defense: April 25, 2017

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: ElBayeh Fréderic Michael

Last First Middle

☑ Master's Thesis ○ Master's Project ○ Doctoral Dissertation

☑ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

☐ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One** ___ **year from the date of submission of my thesis, dissertation or project.**
Two ___ **years from the date of submission of my thesis, dissertation or project.**
Three ___ **years from the date of submission of my thesis, dissertation or project.**

Freddy May 5, 2017

Signature Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

I would first like to thank my thesis adviser Professor Abbas AlHakim of the Department of Mathematics at the American University of Beirut. Professor Abbas was always ready to help me whenever I had a question about my research or writing. He gave me from his own time and considered me a family member rather than a student.

I would also like to thank Professors Nabil Nassif and Stefano Monni, in the Mathematics Department at the American University of Beirut, as second readers for this thesis and I am gratefully indebted for their valuable comments.
Also, I have to thank Professor Wissam Raji for the opportunity he gave to me.

Finally, I would like to thank my mother and my friends for their daily support and continuous encouragement during these two years in AUB.
This work would not have been possible without them.

Not to forget, I would like to thank God from the bottom of my heart, because He was always there for me, leading me and loving me.

Thank you,
Fréderic El Bayeh

An Abstract of the Thesis of

Frederic Michael El Bayeh for Master of Science

Major: Mathematics

Title: A sufficient normality condition for Turing's formula

Given is a multinomial model with infinite number of categories.
For k = 1, 2,... let N_k be the number of categories represented exactly by k observations in the sample; and let p_k be the category probabilities satisfying $0 < p_k < 1$, and $\sum_k p_k = 1$.
In classical statistics, a sample of size n is used to obtain information about the proportions of categories that are observed. The main idea of the current paper is to show how to use the sample to obtain valid information about the categories that were not observed in the sample.
That is, we want to "estimate" the probability:

$$\pi_0 := \sum_{k=1}^{\infty} p_k \mathbb{1}_{[X_k=0]}$$

An "estimator" of the quantity π_0, known as Turing's formula, is given by:

$$T = N_1/n$$

The problem of "estimating" π_0 has many applications: estimating the proportion of new species of animals in a population, studying gene categorization and

discussing data confidentiality.

This thesis establishes a sufficient condition for the asymptotic normality of the non-parametric estimate of π_0 under a fixed distribution $\{p_k\}$ where all $p_k > 0$.

Contents

List of Tables

Chapter 1

Introduction

Consider the population of all the birds in the world and assume that there are infinitely many species in the world enumerated as $k = 1, 2, \cdots$. Also denote the corresponding distribution of proportions $\{p_k; k \geq 1\}$ where p_k is the proportion of the k^{th} bird species in this population satisfying $0 < p_k < 1$ for all k and $\sum_k p_k = 1$.

Suppose a random sample of $n = 2000$ is to be chosen from the population, and let the bird counts for the different species be denoted by $\{X_i; i \geq 1\}$.

For $k = 1, 2, \cdots$, let N_k be the random variable number of species represented exactly k times in the random sample.

For example, if all the 2000 birds come from different species, then $N_1 = 2000$; and if a robin appeared exactly once while all the other species have more than one member, then $N_1 = 1$.

We are interested in estimating p_1, the proportion of birds of species 1 in the population; clearly $\hat{p_1} = X_1/n$ is the maximum likely hood estimator of the parameter p_1.

In the general case, $\hat{p_k} = X_k/n$.

For a sample of size n = 2000 with birds counts given in Table 1.1 and a second version (rearranged in decreasing order of $X_k's$) as shown in Table 1.2.

In this example, p_1 is estimated by $\hat{p_1} = 300/2000 = 0.15$ and p_2 by $\hat{p_2} = 200/2000 = 0.10$ and so on.

The total number of bird species observed in this sample is 30. It is clear that the bird population must have more or equal than just 30 different species.

Here is the question to ask:

What is the total population proportion of birds belonging to species other than those observed in the sample?

This question implies a statistical problem of estimation of the proportion of birds belonging to species that did not appear in the sample. This is a seemingly counter intuitive probability because we are proposing to use a sample to estimate proportions of categories that didn't appear in the sample.

To do so, let us denote this proportion by π_0. That is:

$$\pi_0 = \sum_{k=1}^{\infty} p_k \mathbb{1}_{[X_k=0]}$$

$$where \ \mathbb{1}_{[X_k=0]} = \begin{cases} 1 \ if \ X_k = 0, \\ 0 \ if \ X_k \neq 0 \end{cases}$$

Already defined, we can write:

$$N_k = \sum_{i=1}^{\infty} \mathbb{1}_{[X_i=k]}$$

It is important to notice that π_0 is neither a constant, nor an observable random variable. Also, it is not a statistic since it depends on the unknown proportions of the species not represented in the sample.

Table 1.1: Bird sample

k	1	2	3	4	5	6	7	8	9	10
X_k	300	200	300	200	100	100	100	100	0	100
k	11	12	13	14	15	16	17	18	19	20
X_k	100	80	70	0	30	50	6	1	2	1
k	21	22	23	24	25	26	27	28	29	30
X_k	1	1	0	0	1	1	1	1	1	1
k	31	32	33	34	35	36	37	38	39	...
X_k	50	100	1	1	0	0	0	0	0	

Table 1.2: Rearranged bird sample

k	1	3	2	4	5	6	7	8	10	11
X_k	300	300	200	200	100	100	100	100	100	100
k	32	12	13	16	31	15	17	19	18	20
X_k	100	80	70	50	50	30	6	2	1	1
k	21	22	25	26	27	28	29	30	33	34
X_k	1	1	1	1	1	1	1	1	1	1
k	9	14	23	24	35	36	37	38	39	...
X_k	0	0	0	0	0	0	0	0	0	

Table 1.3: The number of species represented k times in the sample

k	1	2	3	...	6	...	30	...	50	...	70	...	80	...	100	...	200	...	300
N_k	12	1	0	...	1	...	1	...	2	...	1	...	1	...	7	...	2	...	2

Because of the fact that π_0 is not a parameter in the usual sense, we cannot properly speak of estimation of π_0 but rather of prediction of π_0. However, some authors do refer to this problem as to one of estimation of π_0. We follow this use. What is meant by an estimator of π_0 is an observable random variable $\hat{\pi_0}$ in some way close to π_0, denoted here in this paper by T. The quantity π_0, often known as the non coverage probability is interpreted as the probability of discovering a new species i.e the chance that the next bird is of a new or unobserved species. The most essential idea of this thesis is that π_0 can be estimated by the sample using Turing's formula, known also by Good-Turing's formula which was introduced by Good in 1953 but lately credited to Alan Turing.
Turing's formula is given by:

$$T = \frac{N_1}{n}$$

We use the number of species that appeared once in the sample to estimate the proportion of species that didn't appear, which implies:

$$\pi_0 \approx T = \frac{N_1}{n}$$

Clearly in the previous example, $T = 12/2000 = 0.006$ where N_1 is the number of species that appeared only once in the sample.

The problem of estimating a probability of unobserved species may be encountered in several fields such as population biology, species recognition, risk management, discussing data confidentiality.

It is more customary to work with the coverage probability defined by $C = 1 - \pi_0$, and its estimate is $C' = 1 - T$.
Esty [15] gave a sufficient condition for the normality of a $\sqrt{n}$-normalized cov-

erage estimate especially when the behavior of the coverage estimate under an infinite dimensional p_k was discussed. Esty establishes a $\sqrt{n}$-normality law for $C' - C$ where $C = 1 - \pi_0$ and its estimate $C' = 1 - T$, that is:
$\sqrt{n}(C - C')[(N_1/n) + (2N_2/n) - (N_1/n)^2]^{-1/2}$ which converges in distribution to a standard normal.
Unfortunately, Esty's normality law was established not for a fixed $\{p_k\}$ but for a distribution which is allowed to vary as n increases.
If $\{p_k\}$ is fixed, the sufficient condition of Esty never holds and therefore the $\sqrt{n}$-normalized coverage estimate necessarily degenerates at 0.
To straighten out this issue, this paper establishes a sufficient condition for the asymptotic normality of the non-parametric sample coverage estimate under a fixed $\{p_k\}$ but with a normalizing factor $g(n)$ that increases faster than $\sqrt{n}$.

In this thesis, we will:

- Prove that the condition of an earlier limit theorem for T is not satisfied by any particular distribution.

- State and prove a limit theorem for T with a normalizing factor $g(n)$ that increases faster than $\sqrt{n}$, the usual factor in the central limit theorem.

- Show that the conditions of the proposed limit theorem are satisfiable.

- Show how to use the main result to make statistical inference, including constructing confidence intervals and hypothesis testing.

Chapter 2

A Note on Esty's Normality Law

Let $C' = 1 - \frac{N1}{n}$.

Esty establishes a $\sqrt{n}$-normality law for $C' - C$ where $C = 1 - \pi_0$ and its estimate $C' = 1 - T$, that is:

$\sqrt{n}(C - C')[(N_1/n) + (2N_2/n) - (N_1/n)^2]^{-1/2}$ which converges in distribution to a standard normal, in a way allowing the underlying distribution $\{p_k\}$ to vary within a family $\{\{p_k\}_m : m = 1, \cdots\}$ as the sample size n changes to ensure the following imposed conditions would hold:

$$\text{(a)} \quad E(N_1/n) \longrightarrow c_1, \;\; 0 < c_1 < 1 \quad \text{and} \quad \text{(b)} \quad E(N_2/n) \longrightarrow c_2 \geq 0 \qquad (2.1)$$

where $N_2 = \sum \mathbb{1}_{[X_k=2]}$.

Naturally, one would want to have a limit distribution for a particular underlying distribution $\{p_k\}$. However when the distribution is fixed, Equation (2.1) never holds as the following lemma establishes that fact.

Lemma 2.0.1. *Consider a random sample of size n from a multinomial population with probability $\{p_k\}$, then:*

$$(a) \quad \lim_{n\to\infty} E(N_1/n) = 0 \quad \text{and} \quad (b) \quad \lim_{n\to\infty} E(N_2/n) = 0.$$

Proof. For equation (a) we have

$$\begin{aligned}
E\left(\frac{N_1}{n}\right) &= \frac{1}{n}E(N_1) = \frac{1}{n}E\left(\sum \mathbb{1}_{[X_k=1]}\right) \\
&= \frac{1}{n}\sum_{k=1}^{\infty} E(\mathbb{1}_{[X_k=1]}) = \frac{1}{n}\sum_{k=1}^{\infty} P([X_k=1]) \\
&= \frac{1}{n}\sum_{k=1}^{\infty}\binom{n}{1}p_k(1-p_k)^{n-1} \\
&= \sum_{k=1}^{\infty} p_k(1-p_k)^{n-1}
\end{aligned}$$

Since $p_k(1-p_k)^{n-1} \leq p_k$ and $\sum p_k = 1 < \infty$, by the dominated convergence theorem we get

$$\lim_{n\to\infty} E\left(\frac{N_1}{n}\right) = \lim_{n\to\infty}\sum_{k=1}^{\infty} p_k(1-p_k)^{n-1} = \sum_{k=1}^{\infty}\lim_{n\to\infty} p_k(1-p_k)^{n-1} = 0.$$

For part (b) we have that

$$\begin{aligned}
E\left(\frac{N_2}{n}\right) &= \frac{1}{n}E(N_2) = \frac{1}{n}E\left(\sum \mathbb{1}_{[X_k=2]}\right) \\
&= \frac{1}{n}\sum_{k=1}^{\infty} E(\mathbb{1}_{[X_k=2]}) = \frac{1}{n}\sum_{k=1}^{\infty} P([X_k=2]) \\
&= \frac{1}{n}\sum_{k=1}^{\infty}\binom{n}{2}p_k^2(1-p_k)^{n-2} \\
&= \frac{1}{2}\sum_{k=1}^{\infty}(n-1)p_k^2(1-p_k)^{n-2}
\end{aligned}$$

Letting $f(p) = (n-1)p(1-p)^{n-2}$ then we have by straightforward differentiation,

$$\begin{aligned}
f'(p) &= (n-1)(1-p)^{n-2} - (n-1)(n-2)p(1-p)^{n-3} \\
&= (n-1)(1-p)^{n-3}[p(1-n)+1]
\end{aligned}$$

The function $f(p)$ attains its extremum value at $p = \frac{1}{n-1}$ since

$$f'(p) = 0 \iff p(1-n)+1 = 0 \iff p(1-n) = -1 \iff p = \frac{1}{n-1}$$

and to check that it is a maximum, we calculate the second derivative:

$$f''(p) = (n-1)(1-p)^{n-4}[p(-n^2+5n-4)-2],$$

$$f''\left(\frac{1}{n-1}\right) = -(n-1)\left(\frac{n-2}{n-1}\right)^{n-4}(n+6) < 0,$$

and then

$$(n-1)p_k(1-p_k)^{n-2} \leq f\left(\frac{1}{n-1}\right)$$

and

$$f\left(\frac{1}{n-1}\right) = (n-1)\frac{1}{n-1}\left(1-\frac{1}{n-1}\right)^{n-2} = \left(\frac{n-2}{n-1}\right)^{n-2}$$

and thus

$$(n-1)p_k(1-p_k)^{n-2} \leq \left(\frac{n-2}{n-1}\right)^{n-2} < 1.$$

We multiply by p_k on both sides then

$$(n-1)p_k^2(1-p_k)^{n-2} \leq p_k$$

We then use the dominated convergence theorem to interchange the limit and the sum to get

$$\lim_{n\to\infty} E\left(\frac{N_2}{n}\right) = \frac{1}{2}\sum_{k=1}^{\infty}\lim_{n\to\infty}(n-1)p_k^2(1-p_k)^{n-2} = 0.$$

□

The method of Esty [15] is instructive and we are going to follow it in this thesis closely.

The main point of Esty's method is based on direct computation of the limit of the characteristic function of a normalized coverage estimate.

We denote by $K = \{1, 2, \cdots\}$ the index set for the categories.

The method is supported by two different partitions, denoted by $K = M \cup MC$

and $K = I \cup II$. The first partition is designed to support an exchange of a limit operator and an integral operator. The second partition is designed to control the tail probabilities of $\{p_k\}$ as n increases. All the proofs done in this paper are similar to those done by Esty. However, we establish that Esty's first partition (M and MC) is not necessary. Hence in this thesis, we use the second partition (I and II) which depends on a function, $g(n)$, that replaces the $\sqrt{n}$ factor and, therefore, plays an important role in the relevant proofs.

Chapter 3

Preliminary Results

Even though the result of Esty is not satisfied by any particular distribution, the method of Esty [15] is instructive.

Let $K_1 = \{1\}$ and $K_2 = \{2, \cdots\}$. For any $k \in K = K_1 \cup K_2$, let

$$f_k(x) = \begin{cases} p_k & x = 0, \\ -1/n & x = 1, \cdot \\ 0 & x \geq 2. \end{cases}$$

We have that

$$\begin{aligned} C' - C = \pi_0 - T &= \sum_{k=1}^{\infty} p_k \mathbb{1}_{[X_k=0]} - \frac{1}{n} \sum_{k=1}^{\infty} \mathbb{1}_{[X_k=1]} \\ &= \sum_{k=1}^{\infty} \left(p_k \mathbb{1}_{[X_k=0]} - \frac{1}{n} \mathbb{1}_{[X_k=1]} \right) \\ &= \sum_{k=1}^{\infty} f_k(X_k). \end{aligned}$$

Let $Z = C' - C$. We are interested in the asymptotic behavior of $Zg(n)$, where $g(n)$ is a function of n satisfying

$$g(n) = O(n^{1-2\delta}), \tag{3.1}$$

for some $\delta \in (0, 1/4)$, in terms of the limit of the characteristic function, $E[\exp(isZg(n))]$. To begin with, we note that $Z = Z_1 + Z_2$, where $Z_1 = \sum_{K_1} f_k(X_k)$ and $Z_2 =$

$\sum_{K_2} f_k(X_k)$.

Lemma 3.1 below is a well known lemma that allows us to replace X_k by independent Poisson random variables. Lemma 3.2 is due to Bartlett [20].

Lemma 3.0.1. *Let $\{X_k\}$ be the counts of observations in category k, $k = 1, 2 \ldots$, in an random sample from a multinomial population of parameters $\{p_k\}$, then*

$$P(X_k = x_k; k = 1, 2, \ldots) = P\left(Y_k = x_k; k = 1, \ldots \mid \sum Y_k = n\right),$$

where $\{Y_k\}$ are independent Poisson random variables with mean np_k.

Proof. Let $Y_1, \ldots, Y_k$ be independent random Poisson variables with means $np_1, np_2, \cdots, np_k$ respectively.

$$\begin{aligned}
P\left(Y_k = x_k, k = 1, \ldots \mid \sum Y_k = n\right) &= \frac{P\left(Y_k = x_k, k = 1, \ldots \mid \sum_{i=1}^{\infty} Y_i = n\right)}{P\left(\sum_{i=1}^{\infty} Y_i = n\right)} \\
&= \frac{\prod_{i=1}^{\infty} P(Y_i = x_i)}{P\left(\sum_{i=1}^{\infty} Y_i = n\right)} \\
&= \frac{\prod_{i=1}^{\infty} \frac{(np_i)^{x_i}}{x_i!} e^{-np_i}}{\frac{n^n}{n!} e^{-n}} \\
&= \frac{\frac{n^{\sum_{i=1}^{\infty} x_i}}{\prod_{i=1}^{\infty} x_i!} \cdot e^{-n \sum_{i=1}^{\infty} p_i} \cdot \prod_{i=1}^{\infty} p_i^{x_i}}{\frac{n^n}{n!} e^{-n}} \\
&= \frac{n!}{\prod_{i=1}^{\infty} x_i!} \prod_{i=1}^{\infty} p_i^{x_i} \\
&= P(X_k = x_k; k = 1, \ldots)
\end{aligned}$$

where $\sum_{i=1}^{\infty} Y_i = n$ and $\sum_{i=1}^{\infty} p_i = 1$ and this is because n is finite, so $x_k = 0$ for some $k \geq k_0$.

□

Lemma 3.0.2. *Let (U, V) be a two-dimensional random vector with U integer valued. Then*

$$E[\exp(ivV)|U = n] = \frac{1}{2\pi P(U = n)} \int_{-\pi}^{\pi} E[\exp(iu(U - n) + ivV)]du.$$

By the two lemmas 3.1 and 3.2 where $U = \sum Y_k$ and $V = Zg(n)$, we want to evaluate the characteristic function $E(\exp(isZg(n)))$.
First note that, by Stirling's formula, $(2\pi n)^{1/2} P(\sum Y_k = n) \to 1$. Indeed,
We have

$$E\left(e^{isZg(n)} | \sum Y_k = n\right) = \left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} \int_{-\pi}^{\pi} E\left[e^{iu(k-n)+isZg(n)}\right] du$$

But

$$\sum Y_k - n = \sum Y_k - n \sum p_k = \sum (Y_k - np_k)$$

Then

$$E\left(e^{isZg(n)} | \sum Y_k = n\right) = \left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} \int_{-\pi}^{\pi} E\left[e^{iu\sum(Y_k - np_k)+isZg(n)}\right] du.$$

We know that Y_k has a Poisson distribution with parameter np_k i.e. $Y_k \sim P'(np_k)$ so that $\sum Y_k \sim P'\left(\sum np_k\right)$ and thus $\sum Y_k \sim P'(n)$ since $\sum p_k = 1$.
Then

$$P(\sum Y_k = n) = (n^n e^{-n})/n! \text{ thus } \left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} = \frac{e^n n!}{2\pi n^n}$$

but

$$(2\pi n)^{1/2} P\left(\sum Y_k = n\right) = \frac{\sqrt{2\pi n} n^n e^{-n}}{n!}.$$

We know by the Stirling formula that

$$n! \sim n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi} = n^n e^{-n} \sqrt{2\pi n}$$

Hence

$$(2\pi n)^{1/2} P\left(\sum Y_k = n\right) = \frac{\sqrt{2\pi n}\, n^n e^{-n}}{n!} \sim \frac{n!}{n!} = 1$$

and therefore

$$\left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} = \frac{1}{2\pi \dfrac{n^n}{e^n n!}} = \frac{\sqrt{n}}{\sqrt{2\pi n}\left(\dfrac{\sqrt{2\pi}\sqrt{n}\, n^n e^{-n}}{n!}\right)} = \frac{\sqrt{n}}{\sqrt{2\pi}}$$

so

$$E\left(e^{isZg(n)}\right) = \left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} \int_{-\pi}^{\pi} E\left(e^{iu\sum(Y_k - np_k) + isZg(n)}\right) du$$

We denote

$$H_n(s) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\pi}^{\pi} E\left(e^{iu\sum(Y_k - np_k) + isZg(n)}\right) du$$

We will evaluate the limit of $H_n(s)$ using the change of variables formula $t = u\sqrt{n}$, $dt = (t/u)du$ then

$$\begin{aligned} H_n(s) &= \frac{t}{u\sqrt{2\pi}} \int_{-\pi}^{\pi} E\left(e^{in^{-1/2}t\sum(Y_k - np_k) + isZg(n)}\right) \frac{u}{t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} E\left(e^{in^{-1/2}t\sum(Y_k - np_k) + isZg(n)}\right) dt \end{aligned}$$

but $-\pi < u < \pi$ so that $-\pi\sqrt{n} < u\sqrt{n} < \pi\sqrt{n}$ and therefore $-\pi\sqrt{n} < t < \pi\sqrt{n}$ so finally

$$\begin{aligned} H_n(s) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbb{1}_{[-\pi\sqrt{n} < t < \pi\sqrt{n}]} E\left(e^{in^{-1/2}t\sum(Y_k - np_k) + isZg(n)}\right) dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbb{1}_{[|t| < \pi\sqrt{n}]} E\left(e^{in^{-1/2}t\sum(Y_k - np_k) + isZg(n)}\right) dt. \end{aligned}$$

Our first task is to allow the limit operator to exchange with the integral operator. The key element to support this exchange is Equation (3.3).

$$\lim \int |\bar{h}_{n1}| dt = \int \lim |\bar{h}_{n1}| dt. \tag{3.2}$$

Proof of equation (3.3). Let

$$h_n = \mathbb{1}_{[|t| < \pi\sqrt{n}]} E\left(e^{in^{-1/2}t\sum(Y_k - np_k) + isZg(n)}\right)$$

We recall that $K_1 = \{1\}$ and $K_2 = \{2, \ldots\}$ and $K = K_1 \cup K_2$. So we let

$$\begin{aligned} h_{n_1} &= \mathbb{1}_{[|t| < \pi\sqrt{n}]} E\left(e^{in^{-1/2}t(Y_1 - np_1) + isZg(n)}\right) \\ h_{n_2} &= \mathbb{1}_{[|t| < \pi\sqrt{n}]} E\left(e^{in^{-1/2}t\sum_{K_2}(Y_k - np_k) + isZg(n)}\right) \end{aligned}$$

Clearly $h_n = h_{n_1} h_{n_2}$.

Claim 1. *Prove that $|h_{n_2}| \leq 1$ and thus $|h_n| \leq |h_{n_1}|$.*

Proof of Claim.

$$\begin{aligned}
|h_{n_2}| &= \left|\mathbb{1}_{[|t|<\pi\sqrt{n}]} E\left(e^{in^{-1/2}t\sum_{K_2}(Y_k - np_k)+isZg(n)}\right)\right| \\
&= \left|\mathbb{1}_{[|t|<\pi\sqrt{n}]} \sum e^{in^{-1/2}t\sum_{K_2}(Y_k - np_k)+isZg(n)} P[Y_k = x_k]\right| \\
&\leq \sum \left(\left|e^{in^{-1/2}t\sum_{K_2}(Y_k - np_k)+isZg(n)} P[Y_k = x_k]\right|\right) \\
&= \sum \left(\left|e^{i\left(n^{-1/2}t\sum_{K_2}(Y_k - np_k)+sZg(n)\right)}\right| P[Y_k = x_k]\right) \\
&= \sum P_k = 1
\end{aligned}$$

and $|h_n| = |h_{n_1} h_{n_2}| \leq |h_{n_1}|$ so $|h_{n_2}| \leq 1$. □

On the other hand,

$$E\left(e^{iu(Y_1 - np_1)+isf_1(Y_1)g(n)}\right) = \left(e^{iu(-np_1)+isf_1(0)g(n)}\right) P[Y_1 = 0]$$
$$+\left(e^{iu(1-np_1)+isf_1(1)g(n)}\right) P[Y_1 = 1] + \sum_{j=2}^{\infty} e^{iu(j-np_1)} P[Y_1 = j]$$

and since Y_k is a Poisson distribution with parameter np_k, then

$$P[Y_1 = 0] = e^{-np_1}$$

$$P[Y_1 = 1] = np_1 e^{-np_1}$$

Hence

$$E\left(e^{iu(Y_1 - np_1)+isf_1(Y_1)g(n)}\right) = \left(e^{iu(-np_1)+isp_1 g(n)}\right) e^{-np_1} + \left(e^{iu(1-np_1)-isn^{-1}g(n)}\right) np_1 e^{np_1}$$
$$+\sum_{j=2}^{\infty} e^{iu(j-np_1)} P[Y_1 = j]$$

But we know that

$$\sum_{j=0}^{\infty} e^{iu(j-np_1)} P(Y_1 = j) = e^{iu(-np_1)} e^{-np_1} + e^{iu(1-np_1)} np_1 e^{-np_1} + \sum_{j=2}^{\infty} e^{iu(j-np_1)} P[Y_1 = j]$$

So that we get

$$E\left(e^{iu(Y_1-np_1)+isf_1(Y_1)g(n)}\right) = \sum_{j=0}^{\infty} e^{iu(j-np_1)}P[Y_1=j] - e^{-iu(np_1)}e^{-np_1} - e^{iu(1-np_1)}np_1e^{-np_1}$$
$$+e^{iu(-np_1)+isp_1g(n)}e^{-np_1} + e^{-iu(1-np_1)-isn^{-1}g(n)}np_1e^{-np_1}$$

Also notice

$$\begin{aligned}\sum_{j=0}^{\infty} e^{iu(j-np_1)}P[Y_1=j] &= \sum_{j=0}^{\infty} e^{-iunp_1}e^{iuj}P[Y_1=j] \\ &= e^{-iunp_1}\sum_{j=0}^{\infty} e^{iuj}P[Y_1=j] \\ &= e^{-iunp_1}E(e^{iuY_1}) \\ &= e^{-iunp_1}\left(e^{np_1(\cos u-1)}e^{np_1 i\sin u}\right)\end{aligned}$$

The last line is because

$$E(e^{tY}) = e^{\lambda(e^t-1)} = e^{\lambda(\cos u - 1 + i\sin u)} = e^{np_1(\cos u-1+i\sin u)}$$

and if we let $t \to iu$, we obtain

$$E(e^{iuY}) = e^{np_1((\cos u-1)+i\sin u)}$$

Thus,

$$E\left(e^{iu(Y_1-np_1)+isf_1(Y_1)g(n)}\right) = e^{-iunp_1}e^{np_1(\cos u-1)}e^{inp_1\sin u} - e^{-iu(np_1)}e^{-np_1} - e^{iu(1-np_1)}np_1e^{-np_1}$$
$$+e^{iu(-np_1)+isp_1g(n)}e^{-np_1} + e^{-iu(1-np_1)-isn^{-1}g(n)}np_1e^{-np_1}$$

Hence by the triangle inequality we have that

$$
\begin{aligned}
|h_{n_1}| &= \left|\mathbb{1}_{[|t|<\pi\sqrt{n}]} E\left(e^{in^{-1/2}t(Y_1-np_1)+isZg(n)}\right)\right| \\
&= \mathbb{1}_{[|t|<\pi\sqrt{n}]} \big| e^{-iunp_1} e^{np_1(\cos u-1)} e^{inp_1 \sin u} - e^{-iu(np_1)} e^{-np_1} - e^{iu(1-np_1)} np_1 e^{-np_1} \\
&\quad + e^{iu(-np_1)+isp_1 g(n)} e^{-np_1} + e^{-iu(1-np_1)-isn^{-1}g(n)} np_1 e^{-np_1} \big| \\
&\le \mathbb{1}_{[|t|<\pi\sqrt{n}]} |e^{-iunp_1} e^{np_1(\cos u-1)} e^{inp_1 \sin u}| + |e^{-iu(np_1)} e^{-np_1}| + |e^{iu(1-np_1)} np_1 e^{-np_1}| \\
&\quad + |e^{iu(-np_1)+isp_1 g(n)} e^{-np_1}| + |e^{-iu(1-np_1)-isn^{-1}g(n)} np_1 e^{-np_1}| \\
&\le \mathbb{1}_{[|t|<\pi\sqrt{n}]} \left(e^{np_1(\cos u-1)} + e^{-np_1} + np_1 e^{-np_1} + e^{-np_1} + np_1 e^{-np_1}\right) \\
&= \mathbb{1}_{[|t|<\pi\sqrt{n}]} \left[e^{np_1(\cos u-1)} + 2\left(e^{-np_1} + np_1 e^{-np_1}\right)\right]
\end{aligned}
$$

Therefore

$$\lim_{n\to\infty} |h_{n_1}| = \mathbb{1}_{[|t|<\pi\sqrt{n}]} e^{np_1(\cos u-1)} = \mathbb{1}_{[|t|<\pi\sqrt{n}]} e^{np_1(\cos(tn^{-\frac{1}{2}})-1)} = \bar{h}_{n_1}.$$

We then use the Taylor expansion of $\cos u$ to get

$$\cos u - 1 = -\frac{u^2}{2!} + \frac{u^4}{4!} + \cdots$$

Then replace u by $tn^{-1/2}$ to get

$$
\begin{aligned}
\cos(tn^{-\frac{1}{2}}) - 1 &= -\frac{t^2 n^{-1}}{2!} + \frac{t^4 n^{-2}}{4!} + \cdots, \\
n(\cos(tn^{-\frac{1}{2}}) - 1) &= -\frac{t^2}{2} + \frac{t^4}{24n} + \cdots \quad \xrightarrow{n\to\infty} \quad -\frac{t^2}{2},
\end{aligned}
$$

Therefore

$$\lim_{n\to\infty} \bar{h}_{n_1} = e^{-p_1\frac{t^2}{2}} = \bar{h}_1.$$

Now

$$\bar{h}_{n_1} = \mathbb{1}_{[|t|<\pi\sqrt{n}]} \left(e^{np_1(\cos(tn^{-\frac{1}{2}})-1)} + 2e^{-np_1} + 2np_1 e^{-np_1}\right)$$

Therefore

$$
\begin{aligned}
\int \bar{h}_{n_1} dt &= \int \mathbb{1}_{[|t|<\pi\sqrt{n}]} e^{np_1(\cos(tn^{-\frac{1}{2}})-1)} dt + \int \mathbb{1}_{[|t|<\pi\sqrt{n}]} 2e^{-np_1} dt + \int \mathbb{1}_{[|t|<\pi\sqrt{n}]} 2np_1 e^{-np_1} dt \\
&= \int \mathbb{1}_{[|t|<\pi\sqrt{n}]} e^{np_1(\cos(tn^{-\frac{1}{2}})-1)} dt + 2e^{-np_1}(2\pi\sqrt{n}) + +2np_1 e^{-np_1}(2\pi\sqrt{n})
\end{aligned}
$$

By letting $n \to \infty$ we obtain

$$\lim_{n\to\infty} \int |\bar{h}_{n_1}| dt = \lim_{n\to\infty} \int \mathbb{1}_{[|t|<\pi\sqrt{n}]} e^{np_1(\cos(tn^{-\frac{1}{2}})-1)} dt$$

since $4\pi\sqrt{n}e^{np_1} \xrightarrow{n\to\infty} 0$ and $4\pi n\sqrt{n}p_1 e^{-np_1} \xrightarrow{n\to\infty} 0$. Then by using the same change of variables

$$\lim_{n\to\infty} \int |\bar{h}_{n_1}| dt = \lim_{n\to\infty} \int \mathbb{1}_{[|u|<\pi]} e^{np_1(\cos u-1)} \sqrt{n} du = \lim_{n\to\infty} \int_{-\pi}^{\pi} \sqrt{n} e^{np_1(\cos u-1)} du$$

Letting θ be a constant in $(0, 1/2)$, we divide the interval $(-\pi, \pi)$ into

$$\left(-\pi, -\frac{1}{n^{(1-\theta)/2}}\right) \cup \left(-\frac{1}{n^{(1-\theta)/2}}, \frac{1}{n^{(1-\theta)/2}}\right) \cup \left(\frac{1}{n^{(1-\theta)/2}}, \pi\right),$$

We integrate separately on each interval and take the limit

$$\lim_{n\to\infty} \int \bar{h}_{n_1} dt = \lim_{n\to\infty} \int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n} e^{np_1(\cos u-1)} du + \lim_{n\to\infty} \int_{\frac{1}{n^{(1-\theta)/2}} \le |u| < \pi} \sqrt{n} e^{np_1(\cos u-1)} du$$

Claim 2. *Define*

$$\eta_2 = \int_{\frac{1}{n^{(1-\theta)/2}} \le |u| < \pi} \sqrt{n} e^{np_1(\cos u-1)} du$$

then $\lim_{n\to\infty} \eta_2 = 0$

Proof of claim.

$$\begin{aligned}
\lim_{n\to\infty} \eta_2 &= \lim_{n\to\infty} \int_{\frac{1}{n^{(1-\theta)/2}} < |u| < \pi} \sqrt{n} e^{np_1(\cos u-1)} du \\
&\le \lim_{n\to\infty} \int_{\frac{1}{n^{(1-\theta)/2}} \le |u| < \pi} \sqrt{n} e^{np_1(\cos(\frac{1}{n^{(1-\theta)/2}})-1)} du \\
&\le \lim_{n\to\infty} 2\pi\sqrt{n} e^{-np_1\left(1-\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)\right)}
\end{aligned}$$

Since $\frac{1}{n^{(1-\theta)/2}} < |u| < \pi$ so that $\cos\left(\frac{1}{n^{(1-\theta)/2}}\right) < \cos u < -1$ and $\cos u - 1 < \cos\left(\frac{1}{n^{(1-\theta)/2}}\right) - 1$.

But

$$
\begin{aligned}
1-\cos\left(\frac{1}{n^{(1-\theta)/2}}\right) &= \left(1-\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)\right)\cdot\frac{1+\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)}{1+\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)} \\
&= \frac{1-\cos^2\left(\frac{1}{n^{(1-\theta)/2}}\right)}{1+\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)} \\
&= \frac{\sin^2\left(\frac{1}{n^{(1-\theta)/2}}\right)}{1+\cos\left(\frac{1}{n^{(1-\theta)/2}}\right)} \\
&\sim \frac{1}{2}\left(\frac{1}{n^{(1-\theta)/2}}\right)^2 \\
&\sim \frac{1}{2n^{1-\theta}}
\end{aligned}
$$

then we get $\lim_{n\to\infty}\eta_2 \leq \lim_{n\to\infty} 2\pi\sqrt{n}e^{-np_1\mathcal{O}\left(\frac{1}{n^{1-\theta}}\right)} = 0$ $\square$

Now we compute

$$
\lim_{n\to\infty}\eta_1 \equiv \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n}e^{np_1(\cos u-1)}du.
$$

For u satisfying $|u| < \frac{1}{n^{(1-\theta)/2}}$, consider the Taylor expansion of

$$
\cos u - 1 = -\frac{u^2}{2!}+\frac{u^4}{4!}-\frac{u^6}{6!}+\cdots \leq -\frac{u^2}{2}+(u^4+u^8+\cdots) = -\frac{u^2}{2}+\frac{u^4}{1-u^4}
$$

(a geometric sequence of ratio u^4)

Therefore,

$$\lim_{n\to\infty}\eta_1 = \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{np_1(\cos u-1)}du \le \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{np_1\left(-\frac{u^2}{2}+\frac{u^4}{1-u^4}\right)}du$$

$$\le \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{np_1\left(-\frac{u^2}{2}+\frac{\frac{1}{n^{2-2\theta}}}{1-\frac{1}{n^{2-2\theta}}}\right)}du$$

$$= \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{\left(\frac{-np_1u^2}{2}+np_1\frac{\frac{1}{n^{2-2\theta}}}{1-\frac{1}{n^{2-2\theta}}}\right)}du$$

$$= \lim_{n\to\infty}\left[\left(\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{\frac{-np_1u^2}{2}}du\right)\left(e^{O\left(\frac{1}{n^{1-2\theta}}\right)}\right)\right]$$

$$= \lim_{n\to\infty}\left[\left(\int_{|t|<n^{\theta/2}}e^{\frac{-p_1t^2}{2}}dt\right)\left(e^{O\left(\frac{1}{n^{1-2\theta}}\right)}\right)\right]$$

$$= \int_{-\infty}^{+\infty}e^{\frac{-p_1t^2}{2}}dt.$$

Claim 3.

$$\lim_{n\to\infty}\eta_1 \ge \int_{-\infty}^{+\infty}e^{\frac{-p_1t^2}{2}}dt.$$

Proof. Since $\cos u \ge 1-\frac{u^2}{2}$ for all u satisfying $|u| < \dfrac{1}{n^{(1-\theta)/2}}$ then $\cos u - 1 \ge -u^2/2$ and clearly

$$\lim_{n\to\infty}\eta_1 = \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{np_1(\cos u-1)}du$$

$$\ge \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}\sqrt{n}e^{-np_1\frac{u^2}{2}}du$$

$$= \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\theta)/2}}}e^{-p_1t^2/2}dt$$

$$= \int_{-\infty}^{+\infty}e^{\frac{-p_1t^2}{2}}dt.$$

Which proves the claim. □

The claim tells us that

$$\lim_{n\to\infty}\eta_1 = \int_{-\infty}^{+\infty}e^{\frac{-p_1t^2}{2}}dt.$$

and thus

$$\lim_{n\to\infty}\int \bar{h}_{n_1}dt = \lim_{n\to\infty}\eta_1.$$

But we have already proved that

$$\lim_{n\to\infty}\bar{h}_{n_1} = e^{-p_1t^2/2} \quad \text{and} \quad \int \lim_{n\to\infty}\bar{h}_{n_1} = \int_{-\infty}^{\infty} e^{-p_1t^2/2}.$$

and so

$$\lim_{n\to\infty}\int \bar{h}_{n_1}dt = \int \lim_{n\to\infty}\bar{h}_{n_1}dt$$

This finishes the proof of equation (3.3). □

Lemma 3.0.3. *Let h_n and H_n be as defined in Equation (3.3). Then*

$$\lim H_n = \frac{1}{\sqrt{2\pi}}\int \lim h_n dt.$$

Lemma 3.0.4. *For each k we have that $h_n \sim \prod(B_k + E_k)$, where*

$$\begin{aligned}
B_k &= \exp(-itp_kn^{1/2})[\exp(np_k(\exp(itn^{-1/2})-1))] \\
C_k &= \exp(-itp_kn^{1/2})[\exp(isp_kg(n))-1]\exp(-np_k) \\
D_k &= \exp(-itp_kn^{1/2})\exp(itn^{-1/2})[\exp(-isn^{-1}g(n))-1]np_k\exp(-np_k)
\end{aligned}$$

and $E_k = C_k + D_k$,

Proof. For each k,

$$
\begin{aligned}
E\left(e^{it(j-np_k)n^{-1/2}+isZg(n)}\right) &= \sum_{j=0}^{\infty} e^{it(j-np_k)n^{-1/2}+isZg(n)} \cdot P(Y_k = j) \\
&= e^{it(-np_k)n^{-1/2}+isp_kg(n)} \cdot P(Y_k = 0) + e^{it(1-np_k)n^{-1/2}-isn^{-1}g(n)} \cdot P(Y_k = 1) \\
&\quad + \sum_{j=2}^{\infty} e^{it(j-np_k)n^{-1/2}} \cdot P(Y_k = j) \\
&= e^{-itp_kn^{1/2}} e^{isp_kg(n)} e^{-np_k} + e^{it(1-np_k)n^{-1/2}} e^{-isn^{-1}g(n)} np_k e^{-np_k} \\
&\quad + \sum_{j=2}^{\infty} e^{it(j-np_k)n^{-1/2}} P[Y_k = j] \\
&= e^{-itp_kn^{1/2}} e^{isp_kg(n)} e^{-np_k} + e^{itn^{-1/2}} e^{-itn^{1/2}p_k} e^{-isn^{-1}g(n)} np_k e^{-np_k} \\
&\quad + \sum_{j=2}^{\infty} e^{it(j-np_k)n^{-1/2}} P[Y_k = j] \\
&= e^{-itp_kn^{1/2}} e^{isp_kg(n)} e^{-np_k} - e^{itn^{-1/2}} e^{-itn^{1/2}p_k} e^{-isn^{-1}g(n)} np_k e^{-np_k} + \\
&\quad e^{-itn^{1/2}p_k} e^{-np_k} - e^{it(1-np_k)n^{-1/2}} np_k e^{-np_k} + \sum_{j=0}^{\infty} e^{it(j-np_k)n^{-1/2}} P[Y_k = j] \\
&= \sum_{j=0}^{\infty} e^{it(j-np_k)n^{-1/2}} P[Y_k = j] + \left[e^{-itp_kn^{1/2}+isp_kg(n)} - e^{-itn^{1/2}p_k}\right] e^{-np_k} \\
&\quad + \left[e^{itn^{-1/2}} e^{-itn^{1/2}p_k} e^{-isn^{-1}g(n)} - e^{it(1-np_k)n^{-1/2}}\right] np_k e^{-np_k} \\
&= \sum_{j=0}^{\infty} e^{it(j-np_k)n^{-1/2}} P[Y_k = j] + \left[e^{-itp_kn^{1/2}} \left(e^{isp_kg(n)} - 1\right)\right] e^{-np_k} \\
&\quad + \left[e^{itn^{-1/2}} e^{-itn^{1/2}p_k} \left(e^{-isn^{-1}g(n)} - 1\right)\right] np_k e^{-np_k}
\end{aligned}
$$

But we have that

$$
\begin{aligned}
\sum_{j=0}^{\infty} e^{it(j-np_k)n^{-1/2}} P(Y_k = j) &= \sum_{j=0}^{\infty} e^{itjn^{-1/2}} e^{-itn^{1/2}p_k} P(Y_k = j) \\
&= e^{-itn^{1/2}p_k} \sum_{j=0}^{\infty} e^{-itjn^{1/2}} P(Y_k = j) \\
&= e^{-itn^{1/2}p_k} e^{np_k(e^{itn^{-1/2}}-1)}
\end{aligned}
$$

Hence

$$E\left(e^{it(Y_k-np_k)n^{-1/2}+isZg(n)}\right) = \left(e^{-itp_kn^{1/2}}e^{np_k(e^{itn^{-1/2}}-1)}\right) + \left(e^{-itp_kn^{1/2}}(e^{isp_kg(n)}-1)e^{-np_k}\right) + \left[e^{itn^{-1/2}}e^{-itn^{1/2}p_k}(e^{-isn^{-1}g(n)}-1)np_ke^{-np_k}\right]$$

The lemma follows with:

$$\begin{aligned}
B_k &= \exp(-itp_kn^{1/2})[\exp(np_k(\exp(itn^{-1/2})-1))] \\
C_k &= \exp(-itp_kn^{1/2})[\exp(isp_kg(n))-1]\exp(-np_k) \\
D_k &= \exp(-itp_kn^{1/2})\exp(itn^{-1/2})[\exp(-isn^{-1}g(n))-1]np_k\exp(-np_k)
\end{aligned}$$

□

Recalling that $E_k = C_k + D_k$, we are interested in evaluating $\lim\prod(B_k + E_k)$. The following two lemmas are given by Esty [15].

Lemma 3.0.5. *Let $\{\beta_k\}$ and $\{\epsilon_k\}$ be two sequences of complex numbers, and M_n be a sequence of subsets of K, indexed by n. If the following conditions are true:*

(i) $\prod_{M_n} \beta_k \sim \beta$,

(ii) $\sum_{M_n} \epsilon_k \sim \epsilon$,

(iii) $\beta_k \sim 1$ *uniformly,*

(iv) $\epsilon_k \sim 0$ *uniformly,*

(v) *there exists a constant δ_1 such that* $\sum_{M_n} |\beta_k - 1| \leq \delta_1$ *and*

(vi) *there exists a constant δ_2 such that* $\sum_{M_n} \epsilon_k \leq \delta_2$

then

$$\prod_{M_n}(\beta_k + \epsilon_k) \sim \beta e^{\epsilon}$$

where β and ϵ may also depend on n.

Lemma 3.0.6. *For all $k \in K$,*

$$B_k = \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})].$$

Proof. We have that

$$B_k = e^{-itp_k n^{1/2}} \left[e^{np_k(e^{itn^{-1/2}}-1)} \right]$$
$$= e^{-itp_k n^{1/2} + np_k(e^{itn^{-1/2}}-1)}$$

(using Taylor expansion)

$$= e^{-itp_k n^{1/2} + np_k(\frac{itn^{-1/2}}{1!} - \frac{t^2 n^{-1}}{2!} - \frac{it^3 n^{-3/2}}{3!} + \cdots)}$$

$$= e^{-itp_k n^{1/2} + itp_k n^{1/2} - \frac{t^2}{2}p_k - \frac{it^3 n^{-1/2}}{3!}p_k + \cdots}$$

$$= e^{-\frac{t^2}{2}p_k} + O(t^3 n^{-1/2} p_k).$$

Which proves the lemma. □

The next lemma includes some useful facts.

Lemma 3.0.7. (i) *For any complex number x satisfying $|x| < 1$, $|\ln(1+x)| \leq |x|/(1-|x|)$.*

(ii) *For any real number $x \in [0,1)$, $1 - x \geq \exp(-x/(1-x))$.*

(iii) *For any real number $x \in (0, 1/2)$, $1/(1-x) < 1 + 2x$.*

Proof. For part (i), by Taylor's theorem we have

$$|ln(1+x)| = \left| \sum_{j=1}^{\infty} (-1)^{j+1} \frac{x^j}{j} \right| \leq \sum_{j=1}^{\infty} |x|^j = \frac{|x|}{1-|x|}$$

and since $j \geq 1$ then $\frac{1}{j} \leq 1$.

For part (ii), let

$$f(t) = \frac{e^t}{1+t}, \quad t \in [0, \infty)$$

Calculate the derivative of f:

$$f'(t) = \frac{te^t}{(1+t)^2} > 0$$

then $f(t)$ is an increasing function on $[0, \infty)$.

At $t = 0$, $f(t)$ attains its maximal value of 1 since $f(0) = e^0/(1+0) = 1$, then $f(t) \geq f(0)$ and therefore $\frac{e^t}{1+t} \geq 1$. By changing the variables $t = x/(1-x)$ then

$$\frac{1}{1+\frac{x}{1-x}} e^{\frac{x}{1-x}} \geq 1 \implies 1 - x \geq e^{\frac{-x}{1-x}}$$

For part (iii), for any real number $x \in (0, 1/2)$ calculate

$$\frac{1}{1-x} - (1+2x) = \frac{x(-1+2x)}{1-x}$$

since $x \in (0, 1/2)$ then $2x - 1 < 0$ and $x > 0$ and $1 - x > 0$ then

$$\frac{x(2x-1)}{1-x} < 0 \quad \text{hence} \quad \frac{1}{1-x} < 1 + 2x.$$

□

Let us consider a partition of the index set $K = I \cup II$ where

$$I = \left\{k;\ p_k g(n) \leq n^{-\delta}\right\} \quad \text{and } II = \left\{k;\ p_k g(n) > n^{-\delta}\right\}$$

where δ is as in equation (3.1).

Lemma 3.0.8. *$\sum_{II} |E_k| \to 0$ and $\prod_{II}(B_k + E_k)/\prod_{II} B_k \to 1$.*

Proof. We have that

$$E_k = e^{-itp_k n^{1/2}}(e^{isp_k g(n)} - 1)e^{-np_k} + e^{-itp_k n^{1/2}} e^{itn^{-1/2}}(e^{-isn^{-1}g(n)} - 1)e^{-np_k} np_k$$

By the triangle inequality, we get

$$\begin{aligned} |E_k| &= |e^{isp_k g(n)} - 1|e^{-np_k} + |e^{-isn^{-1}g(n)} - 1|e^{-np_k} np_k \\ &\leq (|e^{isp_k g(n)}| + 1)e^{-np_k} + (|e^{-isn^{-1}g(n)}| + 1)e^{-np_k} np_k \\ &= 2e^{-np_k} + 2np_k e^{-np_k} \end{aligned}$$

Therefore

$$\sum_{II} |E_k| \leq 2 \sum_{II} \left(e^{-np_k} + np_k e^{-np_k} \right).$$

Let $f(p_k) = e^{-np_k} + np_k e^{-np_k}$.

The derivative of f is:

$$f'(p_k) = -ne^{-np_k} + ne^{-np_k} - n^2 p_k e^{-np_k} = -n^2 p_k e^{-np_k}$$

For all $k \in II$, $f'(p_k) < 0$ in $(0, 1)$ and the function $f(p_k)$ attains its maximum $p_k = \frac{n^{-\delta}}{g(n)}$ with value

$$\begin{aligned}
f\left(\frac{1}{n^\delta g(n)}\right) &= e^{-n\frac{1}{n^\delta g(n)}} + n\frac{1}{n^\delta g(n)} e^{-n\frac{1}{n^\delta g(n)}} \\
&= e^{-\frac{1}{n^{\delta-1} g(n)}} + \frac{1}{n^{\delta-1} g(n)} e^{-\frac{1}{n^{\delta-1} g(n)}} \\
&= e^{-\frac{n^{1-\delta}}{g(n)}} + \frac{n^{1-\delta}}{g(n)} e^{-\frac{n^{1-\delta}}{g(n)}} \\
&= e^{-\frac{n^{1-\delta}}{g(n)}} \left(1 + \frac{n^{1-\delta}}{g(n)}\right)
\end{aligned}$$

The total number of indices in II is less than or equal to $g(n)n^\delta$ since

$$\#\left\{k : p_k > \frac{n^{-\delta}}{g(n)}\right\} = \#\left\{k : \frac{1}{p_k} < g(n)n^\delta\right\}$$

but p_k is a pmf therefore

$$p_k < \frac{1}{k} \iff k < \frac{1}{p_k} \iff k < g(n)n^\delta,$$

because if $p_k > 1/k$ then $\sum p_k > \sum(1/k)$ but we know that $\sum p_k = 1 < \infty$ and $\sum(1/k)$ is divergent which is not true. Therefore,

$$\begin{aligned}
& f(p_k) \leq f\left(\frac{1}{n^\delta g(n)}\right) \\
\iff & 2\sum_{II} f(p_k) \leq 2\sum_{II} \left(e^{-\frac{n^{1-\delta}}{g(n)}} \left(1 + \frac{n^{1-\delta}}{g(n)}\right)\right) \\
\iff & \sum_{II} |E_k| \leq 2\left[g(n)n^\delta\right] \left(e^{-\frac{n^{1-\delta}}{g(n)}} \left(1 + \frac{n^{1-\delta}}{g(n)}\right)\right)
\end{aligned}$$

where $g(n)n^\delta$ is the number of indices in II and thus

$$\begin{aligned}\sum_{II}|E_k| &\le 2g(n)n^\delta e^{-\frac{n^{1-\delta}}{g(n)}} + 2ne^{-\frac{n^{1-\delta}}{g(n)}} = 2e^{-\frac{n^{1-\delta}}{g(n)}}\Big(g(n)n^\delta + n\Big)\\ &\le 2e^{-\frac{n^{1-\delta}}{O(n^{1-2\delta})}}\Big(O(n^{1-2\delta})n^\delta + n\Big)\\ &\le 2e^{-O(n^\delta)}O(n) \xrightarrow{n\to\infty} 0\end{aligned}$$

Next, it is required to prove that

$$\frac{\prod_{II}(B_k+E_k)}{\prod_{II}B_k} \longrightarrow 1.$$

We let

$$F_k = \frac{\prod_{II}(B_k+E_k)}{\prod_{II}B_k}$$

We evaluate:

$$\begin{aligned}\log F_k &= \log\left(\frac{\prod_{II}(B_k+E_k)}{\prod_{II}B_k}\right)\\ &= \log\left(\prod_{II}(B_k+E_k)\right) - \log\left(\prod_{II}B_k\right)\\ &= \sum_{II}\log(B_k+E_k) - \sum_{II}\log B_k\\ &= \sum_{II}\left(\log(B_k+E_k) - \log B_k\right)\\ &= \sum_{II}\log\left(\frac{B_k+E_k}{B_k}\right)\\ &= \sum_{II}\log\left(1+\frac{E_k}{B_k}\right).\end{aligned}$$

Therefore

$$|\log F_k| \le \left|\sum_{II}\log\left(1+\frac{E_k}{B_k}\right)\right| \le \sum_{II}\left|\log\left(1+\frac{E_k}{B_k}\right)\right|,$$

since $|B_k|$ is bounded away from zero and by the fact that $\lim|E_k| = 0$ and hence by the fact that $\lim|E_k|/|B_k| = 0$ and by applying the first part of lemma 3.7

with $x = E_k/B_k$, we get that

$$|\log F_k| \leq \sum_{II} \left|\log\left(1+\frac{E_k}{B_k}\right)\right| \leq \sum_{II}\left(\frac{\frac{|E_k|}{B_k}}{1-\frac{E_k}{B_k}}\right) = \sum_{II}\left(\frac{\frac{|E_k|}{|B_k|}}{\frac{|B_k|-|E_k|}{|B_k|}}\right)$$
$$= \sum_{II}\left(\frac{|E_k|}{|B_k|-|E_k|}\right) = O\left(\sum_{II}|E_k|\right) \to 0$$

Then $\log F_k \to 0$ so that $F_k \to 1$ and finally

$$\frac{\prod_{II}(B_k+E_k)}{\prod_{II} B_k} \longrightarrow 1.$$

$\square$

Now let us use the condition under which many of the subsequent results are established.

CONDITION 3.0.1. *As $n \to \infty$*

(1) $\sum (g^2(n)/n) p_k e^{-np_k} \to c_1 \geq 0,$

(2) $\sum g^2(n) p_k^2 e^{-np_k} \to c_2 \geq 0,$ *and*

(3) $c_1 + c_2 > 0.$

Lemma 3.0.9. *Under Condition 3.1, all of the conditions of Lemma 3.5 are satisfied with $M_n = I$, $\beta_k = B_k$, $\beta = B$, $\epsilon_k = E_k$ and $\epsilon = E$.*

Proof. We need to check that all six conditions in Lemma 3.5 are satisfied.

(iii) is true because $B_k = e^{-p_k t^2/2 + O(t^3 p_k n^{-1/2})} = e^{-t^2/2 p_k} e^{O(\frac{t^3}{\sqrt{n}} p_k)}$ since p_k is uniformly bounded by $1/(g(n)n^\delta)$ and $p_k/\sqrt{n}$ is uniformly bounded by $1/(g(n)\sqrt{n}n^\delta)$. Therefore, $B_k \to 1$ uniformly as $n \to \infty$.

(i) is true because

$$\prod_I B_k = \prod_I \left(e^{-\frac{t^2}{2}p_k} e^{O(\frac{t^3}{\sqrt{n}}p_k)}\right) = e^{-\frac{t^2}{2}\sum_I p_k} e^{O(\frac{t^3}{\sqrt{n}}\sum_I p_k)}$$

but $\sum_I p_k \leq \sum_{k=k_0}^{\infty} P_k \to 0$ (part of the tail of a convergent series) then $\prod_I B_k \to 1$. Hence, B=1.

(v) is true. Indeed,

$$B_k = \exp\left(-\frac{t^2}{2}p_k + O(t^3 p_k n^{-1/2})\right) \quad \text{and} \quad \frac{-t^2}{2}p_k + O(t^3 p_k n^{-1/2}) \longrightarrow 0,$$

uniformly. Using the Taylor expansion of the exponential we get $e^x - 1 \leq x + x^2 + x^3 + \cdots$ so that

$$|e^x - 1| \leq \frac{|x|}{1-|x|} \quad \text{for} \quad |x| < 1.$$

Here, we take $x = -\frac{t^2}{2}p_k + O(t^3 p_k n^{-1/2})$. Clearly $|x| < 1$ and thus

$$|B_k - 1| \leq \frac{|-\frac{t^2}{2}p_k + O(t^3 p_k n^{-1/2})|}{1 - |-\frac{t^2}{2}p_k + O(t^3 p_k n^{-1/2})|} = O\left(\left|-\frac{t^2}{2}p_k + (t^3 p_k n^{-1/2})\right|\right),$$

and hence

$$\begin{aligned}\sum_I |B_k - 1| &\leq \sum_I O\left(\left|-\frac{t^2}{2}p_k + (t^3 p_k n^{-1/2})\right|\right) = O\left(\frac{t^2}{2}\sum_I p_k + |t|^3 n^{-1/2}\sum_I p_k\right) \\ &< O(t^2 + |t^3|),\end{aligned}$$

then $\delta_1 = O(t^2 + |t^3|)$.

For (ii), (iv) and (vi) we have

$$\begin{aligned}E_k &= C_k + D_k \\ &= e^{-itp_k n^{1/2}}\left(e^{isp_k g(n)} - 1\right)e^{-np_k} + e^{-itp_k n^{1/2}} e^{itn^{-1/2}}\left(e^{-isn^{-1}g(n)} - 1\right)np_k e^{-np_k} \\ &= e^{-np_k} e^{-itp_k\sqrt{n}}\left[\left(e^{isp_k g(n)} - 1\right) + np_k e^{itn^{-1/2}}\left(e^{-isn^{-1}g(n)} - 1\right)\right].\end{aligned}$$

Using the Taylor expansion: $e^{ix}-1 = ix - x^2/2 + O(x^3)$ then

$$
\begin{aligned}
E_k &= e^{-np_k}e^{-itp_k\sqrt{n}}\Bigg[isg(n)p_k - \frac{s^2g^2(n)p_k^2}{2} + O(s^3g^3(n)p_k^3) \\
&\quad + np_k\Big(1+\frac{it}{\sqrt{n}} - \frac{t^2}{2n} + O\Big(\frac{t^3}{n^{3/2}}\Big)\Big)\Big(-\frac{isg(n)}{n} - \frac{s^2g^2(n)}{2n^2} + O\Big(\frac{s^3g^3(n)}{n^3}\Big)\Big)\Bigg] \\
&= e^{-np_k}e^{-itp_k\sqrt{n}}\Bigg[isg(n)p_k - \frac{s^2}{2}g^2(n)p_k^2 + O(s^3g^3(n)p_k^3) \\
&\quad + \Big(np_k + itp_k\sqrt{n} - \frac{t^2}{2}p_k + np_kO\Big(\frac{t^3}{n^{3/2}}\Big)\Big)\Big(-\frac{isg(n)}{n} - \frac{s^2g^2(n)}{2n^2} + O\Big(\frac{s^3g^3(n)}{n^3}\Big)\Big)\Bigg] \\
&= e^{-np_k}e^{-itp_k\sqrt{n}}\Bigg[isg(n)p_k - \frac{s^2}{2}g^2(n)p_k^2 + O(s^3g^3(n)p_k^3) - isg(n)p_k - \frac{s^2}{2}\Big(\frac{g^2(n)p_k}{n}\Big) \\
&\quad + np_kO\Big(\frac{s^3g^3(n)}{n^3}\Big) + st\frac{g(n)}{\sqrt{n}}p_k - \frac{is^2t}{2n^{3/2}}g^2(n)p_k + itp_k\sqrt{n}O\Big(\frac{s^3g^3(n)}{n^3}\Big) + \frac{ist^2}{2}\frac{g(n)}{n}p_k \\
&\quad + \frac{s^2t^2}{4}\frac{g^2(n)}{n^2}p_k - \frac{t^2}{2}p_kO\Big(\frac{s^3g^3(n)}{n^3}\Big) - isg(n)p_kO\Big(\frac{t^3}{n^{3/2}}\Big) - \frac{s^2}{2}\frac{g^2(n)}{n}p_kO\Big(\frac{t^3}{n^{3/2}}\Big) \\
&\quad + np_kO\Big(\frac{t^3}{n^{3/2}}\Big)O\Big(\frac{s^3g^3(n)}{n^3}\Big)\Bigg] \\
&= e^{-np_k}e^{-itp_k\sqrt{n}}\Bigg[-\frac{s^2}{2}g^2(n)p_k^2 - \frac{s^2}{2}\frac{g^2(n)}{n}p_k + st\frac{g(n)}{\sqrt{n}}p_k + \frac{s^2t^2}{4}\frac{g^2(n)}{n^2}p_k - \frac{is^2t}{2n^{3/2}}g^2(n)p_k \\
&\quad + \frac{ist^2}{2}\frac{g(n)}{n}p_k + O\Big(s^3g^3(n)p_k^3\Big) + O\Big(s^3\frac{g^3(n)}{n^2}p_k\Big) + iO\Big(ts^3\frac{g^3(n)p_k}{n^{5/2}}\Big) - O\Big(\frac{s^3t^2}{2}\frac{g^3(n)}{n^3}p_k\Big) \\
&\quad - iO\Big(st^3\frac{g(n)}{n^{3/2}}p_k\Big) - O\Big(\frac{s^2t^3}{2}\frac{g^2(n)}{n^{5/2}}p_k\Big) + O\Big(\frac{s^3t^3}{2}\frac{g^3(n)}{n^{7/2}}p_k\Big)\Bigg] \qquad (A2)
\end{aligned}
$$

We observe the following

(a) For all $k \in I$, $e^{-itp_k\sqrt{n}} \to 1$ uniformly since

$$p_k \le \frac{1}{g(n)n^\delta} \iff p_k\sqrt{n} \le \frac{\sqrt{n}}{g(n)n^\delta} \xrightarrow{n\to\infty} 0, \quad \text{for all } k \in I.$$

(b) Every additive term of E_k converges to 0 uniformly for all $k \in I$, therefore (iv) is checked.

(c) For every term within the brackets in Equation (A_2) denoted by $\mathscr{T}(s,t,n,p_k)$, except the first two terms

$$\sum_I e^{-np_k}|\mathscr{T}(s,t,n,p_k)| \leq \sum e^{-np_k}|\mathscr{T}(s,t,n,p_k)| \to 0$$

uniformly by condition 3.9. The uniform convergence of $\sum_I e^{-np_k}g^2(n)p_k^2$ and $\sum_I e^{-np_k}\frac{g^2(n)}{n}p_k$ are directly guaranteed by condition 3.9 since

$$\sum_I e^{-np_k}g^2(n)p_k^2 \leq \sum_I e^{-np_k}g^2(n)p_k\frac{1}{n} \leq \sum_I e^{-np_k}\frac{g^2(n)}{n}p_k \xrightarrow{n\to\infty} c_1 \geq 0$$

since $p_k \leq \frac{1}{n}$ for all k and

$$\sum_I e^{-np_k}\frac{g^2(n)}{n}p_k \xrightarrow{n\to\infty} c_2 \geq 0$$

then clearly $\sum E_k \to -\frac{s^2}{2}(c_1+c_2) = E$. Therefore (ii) is checked and the uniformity of the convergence for $\sum_I |E_k|$ and hence for $\sum_I |E_k|$ guarantees (vi).

□

Remark 3.0.1. *It may be interesting to note that the third term within the brackets in Equation (A2), $st\frac{g(n)}{\sqrt{n}}p_k$ also satisfies $\sum_I e^{-np_k}st\frac{g(n)}{\sqrt{n}}p_k \to 0$. However, if $g(n) = \sqrt{n}$ as in Esty, this term does not vanish and as a result shows up as an extra term in the asymptotic variance of the normalized coverage estimator in Esty's results.*

Lemma 3.5 and 3.9 give immediately the following corollary.

Corollary 3.0.1. *Under Condition 3.1, $\prod_I(B_k+E_k) \sim \prod_I(B_k\exp(\sum_I E_k))$.*

Lemma 3.0.10. *Under Condition 3.9, $\prod(B_k+E_k) \to Be^E$ where $B = \lim\prod B_k$ and $E = \lim\sum E_k$.*

Proof. $\prod(B_k + E_k) = \prod_I(B_k + E_k)\prod_{II}(B_k + E_k)$.
Using Lemma 3.8 (b), $\prod(E_k + B_k) \sim \prod_{II} B_k$ then

$$\begin{aligned}
\prod(B_k + E_k) &= \prod_I (B_k + E_k) \prod_{II} B_k \\
&\quad \text{(using Lemma 0.3.9)} \\
&\sim \prod_I B_k \left(e^{\sum_I E_k}\right) \prod_{II} B_k \\
&\sim \prod B_k \left(e^{\sum_I E_k}\right) \\
&\quad \text{(using the fact that } \sum_{II} |E_k| \longrightarrow 0 \text{ by Lemma 3.8)} \\
&\sim \prod B_k e^{\sum E_k}
\end{aligned}$$

□

Remark 3.0.2. *At this point, one may see the reason why it is imposed that $g(n) = O(n^{1-2\delta})$ for some small positive δ. If $g(n)$ is allowed to be a sequence increasing to infinity in the order of n or faster, $\sum_{II} E_k \to 0$ cannot be established using the current method. The proof for (a) of Lemma 3.8 will break down. Consequently, the partition $K = I \cup II$ will not effectively support the subsequent proofs.*

Chapter 4

Main Results

Theorem 4.0.1. *Let $g(n)$ be as in Equation (2.1). Under Condition 3.1,*

$$g(n)(C' - C) \xrightarrow{D} N(0, c_1 + c_2).$$

Proof. To prove that $g(n)(C' - C) \xrightarrow{D} N(0, c_1 + c_2)$, we use the Leẃy's Continuity Theorem, i.e. we prove that the characteristic function H_n of $g(n)(C - C')$ converges to the characteristic function of $N(0, c_1 + c_2)$.

We know that $H_n(S) = E(e^{isZg(n)})$ where $Z = C - C'$. We will use Lemma 3.4

$$\lim_{n\to\infty} H_n = \frac{1}{\sqrt{2\pi}} \int \lim_{n\to\infty} h_n dt,$$

and $h_n \sim \prod(B_k + E_k)$ then

$$\lim_{n\to\infty} h_n = \lim_{n\to\infty} \prod(B_k + E_k) = \lim_{n\to\infty} (\prod B_k) e^{\lim_{n\to\infty} \sum E_k}.$$

First let's find $\lim_{n\to\infty} \prod B_k$:

$$\lim_{n\to\infty} \prod B_k = \lim_{n\to\infty} \prod e^{-\frac{t^2}{2}p_k + O(t^3 p_k n^{-1/2})} = \lim_{n\to\infty} e^{-\frac{t^2}{2}\sum p_k + O(t^3 \sum p_k n^{-1/2})} = e^{-t^2/2}$$

Then let's find the limit of the $\sum E_k$:

$$\begin{aligned}\lim_{n\to\infty} \sum E_k &= \lim_{n\to\infty} \sum \left(e^{-np_k} e^{-itp_k\sqrt{n}} \left(-\frac{s^2}{2} g^2(n) p_k^2 - \frac{s^2}{2} \frac{g^2(n)}{n} p_k \right) \right) \\ &= -\frac{s^2}{2} \left(\lim_{n\to\infty} \sum \frac{g^2(n)}{n} p_k e^{-np_k} + \lim_{n\to\infty} \sum g^2(n) p_k e^{-np_k} \right)\end{aligned}$$

Therefore

$$\begin{aligned}
\lim_{n\to\infty} H_n &= \frac{1}{\sqrt{2\pi}} \int \lim_{n\to\infty} h_n dt \\
&= \left(\frac{1}{\sqrt{2\pi}} \int e^{-\frac{t^2}{2}} dt\right) e^{-\frac{s^2}{2}\left(\lim_{n\to\infty} \sum \frac{g^2(n)}{n} p_k e^{-np_k} + \lim_{n\to\infty} \sum g^2(n) p_k e^{-np_k}\right)} \\
&= e^{-\frac{s^2}{2}(c_1+c_2)}
\end{aligned}$$

Because $(\frac{1}{\sqrt{2\pi}} \int e^{-\frac{t^2}{2}} dt = 1$ and by Condition 3.1

$$\sum \frac{g^2(n)}{n} p_k e^{-np_k} \longrightarrow c_1 \geq 0$$

$$\sum g^2(n) p_k e^{-np_k} \longrightarrow c_2 \geq 0$$

as $n \to \infty$.

Clearly, the limit of H_n is the characteristic function of a normal distribution with mean 0 and variance $\sigma^2 = c_1 + c_2 > 0$.

By the Leǘy's continuity theorem

$$Zg(n) \xrightarrow{d} N(0, c_1 + c_2)$$

$$g(n)(C' - C) \xrightarrow{d} N(0, c_1 + c_2)$$

$\square$

Given a $g(n)$ satisfying Equation (3.1), Condition (3.1) imposes a rate of convergence of $\{p_k\}$. To see that and the condition of Theorem 4.1 describes a non-empty class of distribution, we consider the following example.

Example 4.1 Let

$$p_k = \frac{2}{(k+1)^2}, \quad k = 1, 2, \ldots$$

Condition 3.1 holds if and only if $g(n) = O(n^{3/4})$. To see this, we have

(1)

$$
\begin{aligned}
&\frac{g^2(n)}{n}\int_1^{\infty}\frac{2}{(x+1)^2}e^{-\frac{2n}{(x+1)^2}}dx \\
&= -2\frac{g^2(n)}{n}\int_{1/2}^{0}e^{-2nt^2}dt \quad (\text{use the change of variables } t=\frac{1}{x+1}) \\
&= 2\frac{g^2(n)}{n}\int_0^{1/2}e^{-2nt^2}dt \\
&= 2\frac{g^2(n)}{n}\frac{1}{2\sqrt{n}}\int_0^{\sqrt{n}}e^{-\frac{t^2}{2}}dt \quad (\text{use the change of variables } u=2\sqrt{n}t \text{ therefore } du=2\sqrt{n}dt) \\
&= 2\frac{g^2(n)}{n}\frac{1}{2\sqrt{n}}\frac{\sqrt{2\pi}}{\sqrt{2\pi}}\int_0^{\sqrt{n}}e^{-\frac{t^2}{2}}dt \\
&= O\left(\frac{g^2(n)}{n\sqrt{n}}\right) = O\left(\frac{g^2(n)}{n^{3/2}}\right)
\end{aligned}
$$

since when $n\to\infty$, $\frac{1}{\sqrt{2\pi}}\int_0^{\sqrt{n}}e^{-t^2/2}dt$ goes to a constant but $O\left(\frac{g^2(n)}{n^{3/2}}\right)$ is a non-zero constant when $g^2(n)=kn^{3/2} \iff g(n)=kn^{3/4} \iff g(n)=O(n^{3/4})$.

(2) Similarly using the same change of variables

$$
\begin{aligned}
&g^2(n)\int_1^{\infty}\frac{4}{(x+1)^4}e^{-\frac{2n}{(x+1)^2}}dx \\
&= 4g^2(n)\int_0^{1/2}t^2e^{-2nt^2}dt \\
&= \frac{4g^2(n)}{(2\sqrt{n})^3}\int_0^{\sqrt{n}}t^2e^{-t^2/2}dt \\
&= \frac{4g^2(n)}{(2\sqrt{n})^3}\frac{\sqrt{2\pi}}{\sqrt{2\pi}}\int_0^{\sqrt{n}}t^2e^{-t^2/2}dt \\
&= O\left(\frac{g^2(n)}{n^{3/2}}\right)
\end{aligned}
$$

since $\frac{1}{\sqrt{2\pi}}\int_0^{\sqrt{n}}t^2e^{-t^2/2}dt$ goes to constant when n goes to infinity.
Same as above, $O\left(\frac{g^2(n)}{n^{3/2}}\right)$ goes to a non-zero constant if and only if $g(n)=O(n^{3/4})$.

Remark 4.1 This example proves that there is at least one distribution satisfying Condition 3.1 when the conditions of Esty didn't satisfy an distribution.

Let us consider the following condition.

CONDITION 4.0.1. *As $n \to \infty$,*

(1) $\frac{g^2(n)}{n^2} E(N_1) \to c_1 \geq 0$,

(2) $\frac{g^2(n)}{n^2} E(N_2) \to \frac{c_2}{2} \geq 0$

(3) $c_1 + c_2 > 0$.

Lemma 4.0.2. *Condition 3.1 and Condition 4.1 are equivalent.*

Proof.

$$\begin{aligned}
\frac{g^2(n)}{n^2} E(N_1) &= \frac{g^2(n)}{n^2} E\left(\sum \mathbb{1}_{[X_k=1]}\right) \\
&= \frac{g^2(n)}{n^2} \sum E(\mathbb{1}_{[X_k=1]}) \\
&= \frac{g^2(n)}{n^2} \sum P[X_k = 1] \\
&= \frac{g^2(n)}{n^2} n \sum p_k (1-p_k)^{n-1} \\
&= \frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} + \frac{g^2(n)}{n^2} n \sum_{II} p_k (1-p_k)^{n-1}
\end{aligned}$$

Using the partition $K = I \cup II$ where

$$I = \left\{k : p_k g(n) \leq n^{-\delta}\right\} = \left\{k : p_k \leq \frac{1}{g(n) n^\delta}\right\}$$

and $II = I^C$. Let $f(p) = pe^{-np}$ then $f'(p) = (1-np)e^{-np}$.
Notice that f' is negative on $(\frac{1}{n}, 1]$ i.e. on $(\frac{1}{n^\delta g(n)}, 1]$ for large n so that f is decreasing for large n and then

$$f(p_k) \leq f\left(\frac{1}{n^\delta g(n)}\right) \iff p_k e^{-np_k} \leq \frac{1}{n^\delta g(n)} e^{-n\left(\frac{1}{n^\delta g(n)}\right)},$$

but since $p_k > \frac{n^{-\delta}}{g(n)}$ on I then $-(n-1)p_k \leq \dfrac{-(n-1)}{g(n)n^{\delta}}$ and since $1-x \leq e^{-x}$ then

$$(1-p_k)^{n-1} \leq e^{-(n-1)p_k} \leq e^{-\frac{n-1}{g(n)n^{\delta}}}$$

Therefore

$$\begin{aligned}
\frac{g^2(n)}{n^2} n \sum_{II} p_k(1-p_k)^{n-1} &\leq \frac{g^2(n)}{n^2} n \sum_{II} p_k e^{-(n-1)p_k} \\
&\leq \frac{g^2(n)}{n^2} n \sum_{II} \frac{1}{g(n)n^{\delta}} e^{\frac{-(n-1)}{g(n)n^{\delta}}} \\
&\leq \frac{g^2(n)}{n^2} n(g(n)n^{\delta}) \frac{1}{g(n)n^{\delta}} e^{\frac{-(n-1)}{g(n)n^{\delta}}} \\
&= \frac{g^2(n)}{n} e^{\frac{-(n-1)}{g(n)n^{\delta}}} \\
&= \frac{O(n^{1-2\delta})^2}{n} O\left(e^{\frac{-(n-1)}{n^{1-2\delta}n^{\delta}}}\right) \\
&= O(n^{1-4\delta})O(e^{-n^{\delta}}) \to 0
\end{aligned}$$

as $n \to \infty$, where $g(n)n^{\delta}$ is the maximum number of terms in II. Now by the Sandwich Theorem, since $\frac{g^2(n)}{n^2}E(N_1) \geq 0$ then

$$\lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_{II} p_k(1-p_k)^{n-1} = 0.$$

Now,

$$\frac{g^2(n)}{n^2}E(N_1) = \frac{g^2(n)}{n^2} n \sum_{I} p_k(1-p_k)^{n-1} + \frac{g^2(n)}{n^2} n \sum_{II} p_k(1-p_k)^{n-1},$$

thus we have

$$\lim_{n\to\infty} \frac{g^2(n)}{n^2}E(N_1) = \lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_{I} p_k(1-p_k)^{n-1}.$$

On the other hand, since $-np_k + p_k \leq -np_k + \sup_I(p_k)$
then $e^{-(n-1)p_k} \leq e^{-np_k + \sup_I(p_k)} = e^{\sup_I(p_k)}e^{-np_k}$ and thus

$$\frac{g^2(n)}{n^2} n \sum_{I} p_k(1-p_k)^{n-1} \leq \frac{g^2(n)}{n^2} n \sum_{I} e^{-(n-1)p_k} \leq \frac{g^2(n)}{n^2} n e^{\sup_I p_k} \sum_{I} p_k e^{-np_k}$$

and hence by the definition of I, $\lim_{n\to\infty} e^{\sup_I p_k} = 1$.

Furthermore, by applying part (2) of Lemma 3.7, we have:

$$(1-p_k) \geq e^{-\frac{p_k}{1-p_k}} \iff (1-p_k)^{n-1} \geq e^{-\frac{n-1}{1-p_k}p_k}.$$

then

$$\frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} \geq \frac{g^2(n)}{n^2} n \sum_I p_k e^{-\frac{n-1}{1-p_k}p_k} \geq \frac{g^2(n)}{n^2} n \sum_I p_k e^{-\frac{np_k}{1-\sup_I p_k}}$$

Also by part (3) of Lemma 3.7, we know that

$$\frac{1}{1-\sup_I p_k} < 1 + 2\sup_I p_k \quad \text{then} \quad \frac{-np_k}{1-\sup_I p_k} \geq -np_k(1+2\sup_I p_k)$$

and we also have that

$$p_k \cdot \sup_I p_k \leq (\sup_I p_k)^2 \iff -2np_k(\sup_I p_k) \geq -2n(\sup_I p_k)^2$$

then

$$\begin{aligned}
\frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} &\geq \frac{g^2(n)}{n^2} n \sum_I p_k e^{-np_k(1+2\sup_I p_k)} \\
&\geq \frac{g^2(n)}{n^2} n e^{-2n(\sup_I p_k)^2} \sum_I p_k e^{-np_k} \\
&\geq \frac{g^2(n)}{n^2} n \sum_I p_k e^{-np_k}
\end{aligned}$$

since by the definition of I, $\lim_{n\to\infty} e^{-2n(\sup_I p_k)^2} = 1$. Therefore,

$$\lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} = \lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_I p_k e^{-np_k}.$$

Clearly,

$$\begin{aligned}
\lim_{n\to\infty} \frac{g^2(n)}{n^2} E(N_1) &= \lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} \\
&= \lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum_I p_k e^{-np_k} \\
&= \lim_{n\to\infty} \frac{g^2(n)}{n^2} n \sum p_k e^{-np_k} \\
&= \lim_{n\to\infty} \frac{g^2(n)}{n} \sum p_k e^{-np_k}
\end{aligned}$$

Similarly for (2).

$$
\begin{aligned}
2\frac{g^2(n)}{n^2}E(N_2) &= 2\frac{g^2(n)}{n^2}E\left(\sum \mathbb{1}_{[X_k=2]}\right) = 2\frac{g^2(n)}{n^2}\sum\left(\mathbb{1}_{[X_k=2]}\right) = 2\frac{g^2(n)}{n^2}\sum P[X_k=2] \\
&= \frac{g^2(n)}{n^2}\sum\binom{n}{2}p_k^2(1-p_k)^{n-2} = \frac{g^2(n)}{n^2}n(n-1)\sum p_k^2(1-p_k)^{n-2} \\
&\leq \frac{g^2(n)}{n^2}n^2\sum p_k^2(1-p_k)^{n-2} = g^2(n)\sum p_k^2(1-p_k)^{n-2} \\
&= g^2(n)\sum_{I} p_k^2(1-p_k)^{n-2} + g^2(n)\sum_{II} p_k^2(1-p_k)^{n-2}
\end{aligned}
$$

Let's compute the second term:

$$
\begin{aligned}
g^2(n)\sum_{II} p_k^2(1-p_k)^{n-2} &\leq g^2(n)\sum p_k^2 e^{-(n-2)p_k} \\
&\leq g^2(n)\sum\left(\frac{1}{n^\delta g(n)}\right)^2 e^{-(n-2)}\left(\frac{1}{n^\delta g(n)}\right) \\
&\leq g^2(n)\frac{1}{(n^\delta g(n))^2}(n^\delta g(n))e^{-(n-2)}\left(\frac{1}{n^\delta g(n)}\right) \\
&\leq \frac{O(n^{1-2\delta})}{n^\delta}O\left(e^{-(n-2)}\left(\frac{1}{n^\delta g(n)}\right)\right) \\
&= O(n^{1-3\delta})O(e^{-n^\delta}) \xrightarrow{n\to\infty} 0.
\end{aligned}
$$

Clearly,

$$
2\frac{g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_{II} p_k^2(1-p_k)^{n-2} \xrightarrow{n\to\infty} 0.
$$

By the sandwich theorem, since $2\frac{g^2(n)}{n^2}E(N_2) \geq 0$ then

$$
\lim_{n\to\infty} 2\frac{g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_{II} p_k^2(1-p_k)^{n-2} = 0.
$$

Now

$$
\lim_{n\to\infty} 2\frac{g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_{I} p_k^2(1-p_k)^{n-2} = \lim_{n\to\infty} 2\frac{g^2(n)}{n^2}E(N_2).
$$

and

$$
\begin{aligned}
\frac{2g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_I p_k^2(1-p_k)^{n-2} &\le \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{-(n-2)p_k} \\
&\le \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{-np_k+2\sup_I p_k} \\
(\text{ since } \lim_{n\to\infty} e^{2\sup_I p_k}=1) \qquad &\le \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{-np_k} \\
&\le g^2(n)\sum_I p_k^2 e^{-np_k}.
\end{aligned}
$$

Furthermore by applying part (2) and (3) of Lemma 3.7,

$$
\begin{aligned}
\frac{2g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_I p_k^2(1-p_k)^{n-2} &\ge \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{\frac{-(n-2)p_k}{1-p_k}} \\
&\ge \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{\left(\frac{-np_k}{1-\sup_I p_k}\right)} \\
&\ge \frac{g^2(n)}{n^2}n(n-1)e^{-2n(\sup_I p_k)^2}\sum_I p_k^2 e^{-np_k} \\
(\text{ since } e^{-2n(\sup_I p_k)^2} \xrightarrow{n\to\infty} 1) \quad &\ge \frac{g^2(n)}{n^2}n(n-1)\sum_I p_k^2 e^{-np_k} \\
&\ge g^2(n)\sum_I p_k^2 e^{-np_k},
\end{aligned}
$$

Therefore,

$$
\lim_{n\to\infty}\frac{2g^2(n)}{n^2}\frac{n(n-1)}{2}\sum_I p_k^2(1-p_k)^{n-2} = \lim_{n\to\infty} g^2(n)\sum_I p_k^2 e^{-np_k},
$$

and thus

$$
\lim_{n\to\infty}\frac{2g^2(n)}{n^2}E(N_2) = \lim_{n\to\infty} g^2(n)\sum p_k^2 e^{-np_k},
$$

and the equivalence between Condition 3.1 and 4.1 is established. $\square$

Theorem 4.0.3. *If there is a $g(n)$ satisfying Equation (3.1) and Condition (4.1) then*

$$
\frac{n(C'-C)}{\sqrt{E(N_1)+2E(N_2)}} \xrightarrow{D} N(0,1)
$$

Proof. If $g(n) = O(n^{1-2\delta})$, we want to prove that
$$\frac{n(C'-C)}{\sqrt{E(N_1)+2E(N_2)}} \xrightarrow{D} N(0,1).$$
To see this, let's standardize theorem 4.1:
$$g(n)(C'-C) \xrightarrow{D} N(0, c_1+c_2).$$
where $\mu = 0$ and $\sigma^2 = c_1 + c_2$ then
$$\frac{g(n)(C'-C)-\mu}{\sqrt{c_1+c_2}} \xrightarrow{D} N(0,1) \quad \text{ie} \quad \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}} \xrightarrow{D} N(0,1).$$
On the other hand, multiplying by $g(n)\sqrt{c_1+c_2}$ up and down,
$$\begin{aligned}\frac{n(C'-C)}{\sqrt{E(N_1)+2E(N_2)}} &= \frac{n\sqrt{c_1+c_2}}{g(n)\sqrt{E(N_1)+2E(N_2)}} \cdot \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}} \\ &= \frac{\sqrt{c_1+c_2}}{\sqrt{\frac{g^2(n)}{n^2}E(N_1)+2\frac{g^2(n)}{n^2}E(N_2)}} \cdot \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}}\end{aligned}$$
Now by Condition 4.1, as $n \to \infty$,
$$\sqrt{\frac{g^2(n)}{n^2}E(N_1)+2\frac{g^2(n)}{n^2}E(N_2)} \longrightarrow \sqrt{c_1+c_2}.$$
and hence
$$\frac{\sqrt{c_1+c_2}}{\sqrt{\frac{g^2(n)}{n^2}E(N_1)+2\frac{g^2(n)}{n^2}E(N_2)}} \longrightarrow 1.$$
Now by Slutsky's theorem
$$\frac{n\sqrt{c_1+c_2}}{g(n)\sqrt{E(N_1)+2E(N_2)}} \cdot \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}} \xrightarrow{D} N(0,1).$$
which finishes the proof. □

Remark 4.2 The statement of Theorem 4.2 can be re-written as
$$\frac{\sqrt{n}(C'-C)}{\sqrt{\frac{E(N_1)}{n}+2\frac{E(N_2)}{n}}} \xrightarrow{D} N(0,1).$$
which resembles very much Theorem 4 of Esty except the third term in the variance of Esty is missing. However, it is to be noted that in the current context, the coverage statistic, even though its normalized form can be expressed as above, is not normalized by $\sqrt{n}$ but by $g(n)$ satisfying $g(n)/\sqrt{n} \to \infty$.

As a consequence of Theorem 4.1, we have the following theorem.

Theorem 4.0.4. *If there is a $g(n)$ satisfying Equation (3.1) and Condition 4.1 then*

$$\frac{n(C'-C)}{\sqrt{N_1+2N_2}} \xrightarrow{D} N(0,1).$$

Proof. Let $\hat{c_1}$ and $\hat{c}_2$ be the estimate of c_1 and c_2 respectively.
Let $\hat{c}_1 = \frac{g^2(n)}{n^2}N_1$ and $\hat{c}_2 = 2\frac{g^2(n)}{n^2}N_2$, then

$$\begin{aligned} E(\hat{c}_1) &= E\left(\frac{g^2(n)}{n^2}N_1\right) = \frac{g^2(n)}{n^2}E(N_1) \longrightarrow c_1 \\ E(\hat{c}_2) &= E\left(2\frac{g^2(n)}{n^2}N_2\right) = 2\frac{g^2(n)}{n^2}E(n_2) \longrightarrow c_2 \end{aligned}$$

by Condition 4.1. If suffices to show that $\hat{c}_1$ and $\hat{c}_2$ are consistent estimates of c_1 and c_2 respectively.
Using Markov's inequality:

$$\begin{aligned} P[|\hat{c}_1 - c_1| \geq \epsilon] = P[(\hat{c}_1-c_1)^2 \geq \epsilon^2] &\leq \frac{E[(\hat{c}_1-c_1)^2]}{\epsilon^2} \\ &= \frac{E[(\hat{c}_1-\mu_{\hat{c}_1}+\mu_{\hat{c}_1}-c_1)^2]}{\epsilon^2} \\ &= \frac{E[\hat{c}_1-\mu_{\hat{c}_1}]^2 + E[\mu_{\hat{c}_1}-c_1]^2 + 2E[(\hat{c}_1-\mu_{\hat{c}_1})(\mu_{\hat{c}_1}-c_1)]}{\epsilon^2} \\ &(\text{ since } E[(\hat{c}_1-\mu_{\hat{c}_1})(\mu_{\hat{c}_1}-c_1)] \to 0\ ) \\ &= \frac{V(\hat{c}_1) + (\mu_{\hat{c}_1}-c_1)^2}{\epsilon^2} \end{aligned}$$

and $(\mu_{\hat{c}_1} - c_1)^2 \to 0$ by Condition 4.1. Then to prove the consistency of $\hat{c}_1$ or in other words that $\hat{c}_1 \to c_1$, it suffices to show that $V(\hat{c}_1) = 0$ as $n \to \infty$. Indeed,

$$\begin{aligned}
V(\hat{c}_1) = V\left(\frac{g^2(n)}{n^2}N_1\right) &= \frac{g^4(n)}{n^4}V(N_1) = \frac{g^4(n)}{n^4}COV(N_1, \dot{N}_1) \\
&= \frac{g^4(n)}{n^4}COV\left(\sum_k \mathbb{1}_{[X_k=1]}, \sum_j \mathbb{1}_{[X_j=1]}\right) \\
&= \frac{g^4(n)}{n^4}\sum_k \sum_j COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right) \\
&= \frac{g^4(n)}{n^4}\left(\sum_k COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_k=1]}\right) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right) \\
&= \frac{g^4(n)}{n^4}\left(\sum_k V(N_1) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right) \\
&= \frac{g^4(n)}{n^4}\left(\sum_k \left(E(\mathbb{1}^2_{[X_k=1]}) - E^2(\mathbb{1}_{[X_k=1]})\right) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right) \\
&= \frac{g^4(n)}{n^4}\left(\sum_k E(\mathbb{1}_{[X_k=1]}) - \sum_k E^2(\mathbb{1}_{[X_k=1]}) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right) \\
&= \frac{g^4(n)}{n^4}\left(E\left(\sum_k \mathbb{1}_{[X_k=1]}\right) - \sum_k E^2(\mathbb{1}_{[X_k=1]}) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right) \\
&\leq \frac{g^4(n)}{n^4}\left(E(N_1) + \sum_{k \neq j} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right)
\end{aligned}$$

By the first part of Condition 4.1,

$$\frac{g^4(n)}{n^4}E(N_1) \to 0 \text{ since } \frac{g^2(n)}{n^2} = \frac{O(n^{1-2\delta})^2}{n^2} \sim n^{-4\delta} \to 0 \text{ as } n \to \infty$$

On the other hand,

$$\begin{aligned}
&\frac{g^4(n)}{n^4}\left(\sum_{j\neq k} COV\left(\mathbb{1}_{[X_k=1]},\mathbb{1}_{[X_j=1]}\right)\right)\\
&=\frac{g^4(n)}{n^4}\sum_{j\neq k}\left(E(\mathbb{1}_{[X_j=1]})(\mathbb{1}_{[X_k=1]})-E(\mathbb{1}_{[X_j=1]})E(\mathbb{1}_{[X_k=1]})\right)\\
&=\frac{g^4(n)}{n^4}\sum_{j\neq k}\left(E(\mathbb{1}_{[X_j=1,X_k=1]})-E(\mathbb{1}_{[X_j=1]})E(\mathbb{1}_{[X_k=1]})\right)\\
&=\frac{g^4(n)}{n^4}\sum_{j\neq k}\left(\frac{n!}{1!1!(n-2)!}p_jp_k(1-p_j-p_k)^{n-2}-np_j(1-p_j)^{n-1}np_k(1-p_k)^{n-1}\right)\\
&=\frac{g^4(n)}{n^4}\sum_{j\neq k}(n(n-1)p_jp_k(1-p_j-p_k)^{n-2}-n^2p_jp_k(1-p_j)^{n-1}(1-p_k)^{n-1})\\
&\leq\frac{g^4(n)}{n^4}\sum_{j\neq k}(n^2p_jp_k(1-p_j-p_k)^{n-2}-n^2p_jp_k(1-p_j)^{n-1}(1-p_k)^{n-1})\\
&\leq\frac{g^4(n)}{n^2}\sum_{j\neq k}(p_jp_k(1-p_j-p_k)^{n-2}-p_jp_k(1-p_j)^{n-1}(1-p_k)^{n-1}).
\end{aligned}$$

but

$$(1-p_j-p_k)^{n-2}\leq(1-p_j-p_k+p_jp_k)^{n-2}=(1-p_k)^{n-2}(1-p_j)^{n-2},$$

then

$$
\begin{aligned}
&\frac{g^4(n)}{n^4}\left(\sum_{j\neq k} COV\left(\mathbb{1}_{[X_k=1]}, \mathbb{1}_{[X_j=1]}\right)\right)\\
&\leq \frac{g^4(n)}{n^2}\sum_{j\neq k}\left(p_jp_k(1-p_j)^{n-2}(1-p_k)^{n-2}-p_jp_k(1-p_j)^{n-1}(1-p_k)^{n-1}\right)\\
&= \frac{g^4(n)}{n^2}\sum_{j\neq k}\left(p_jp_k(1-p_j)^{n-2}(1-p_k)^{n-2}(p_k+p_j-p_kp_j)\right)\\
&\leq \frac{g^4(n)}{n^2}\sum_{j\neq k}\left(p_jp_k(1-p_j)^{n-2}(1-p_k)^{n-2}(p_k+p_j)\right)
\end{aligned}
$$

and by symmetry

$$
\begin{aligned}
&\leq 2\frac{g^4(n)}{n^2}\sum_{j,k}(p_j)^2p_k(1-p_j)^{n-2}(1-p_k)^{n-2}\\
&= \frac{2}{n}\left(\frac{g^2(n)}{n}\sum p_k(1-p_k)^{n-2}\right)\left(g^2(n)\sum p_k^2(1-p_k)^{n-2}\right)\\
&= \frac{2}{n}\left(\frac{g^2(n)}{n}E(\frac{N_1}{n})\frac{1}{1-p_k}\right)\left(2\frac{g^2(n)}{n-1}E(\frac{N_2}{n})\right)\\
&\longrightarrow 0 \text{ as } n\to\infty \quad \text{by Lemma 3.1 and Condition 4.1.}
\end{aligned}
$$

Hence $V(\hat{c}_1) \longrightarrow 0$ and $n \to \infty$ and $P[|\hat{c}_1 - c_2|] \geq \epsilon] \longrightarrow 0$ as $n \to \infty$ which is the desired result and therefore $\hat{c}_1$ is consistent.

Similarly, we want to prove that $\hat{c}_2 \to c_2$ as $n \to \infty$. We know that by Markov's:

$$P[|\hat{c}_2 - c_2| \geq \epsilon] \leq \frac{V(\hat{c}_2) + (\mu_{\hat{c}_2} - c_2)^2}{\epsilon^2}$$

but $\mu_{\hat{c}_2} - c_2 \to 0$ by Condition 4.1. It remains to show that $V(\hat{c}_2) \to 0$ as $n \to \infty$

$$
\begin{aligned}
V(\hat{c}_2) &= V\left(2\frac{g^2(n)}{n^2}N_2\right) = 4\frac{g^4(n)}{n^4}V(N_2)\\
&\leq 4\frac{g^4(n)}{n^4}\left(E(N_2) + \sum_{k\neq j} COV\left(\mathbb{1}_{[X_k=2]}, \mathbb{1}_{[X_j=2]}\right)\right).
\end{aligned}
$$

By the same arguments as above, the term converges to 0 since

$$4\frac{g^4(n)}{n^4}E(N_2) = 4\frac{g^4(n)}{n^4}E\left(\sum \mathbb{1}_{[X_k=2]}\right).$$

But we know that

$$\begin{aligned}
E\left(\frac{N_2}{n}\right) &= \frac{1}{2}\sum(n-1)p_k^2(1-p_k)^{n-2}\\
\frac{2}{n}E\left(N_2\right) &= \sum(n-1)p_k^2(1-p_k)^{n-2}\\
\frac{2}{n}E\left(N_2\right) &\leq \sum np_k^2(1-p_k)^{n-2}\\
\frac{2}{n^2}E\left(N_2\right) &\leq \sum p_k^2(1-p_k)^{n-2}
\end{aligned}$$

then

$$\begin{aligned}
4\frac{g^4(n)}{n^4}E(N_2) = 2\frac{g^4(n)}{n^2}\left(\frac{2}{n^2}E(N_2)\right) &\leq 2\frac{g^4(n)}{n^2}\left(\sum p_k^2(1-p_k)^{n-2}\right)\\
&= 2\frac{g^2(n)}{n^2}\left(g^2(n)\sum p_k^2(1-p_k)^{n-2}\right) \xrightarrow{n\to\infty} 0.
\end{aligned}$$

For the second term, let's verify the inequality first:

$$1-(1-p_k)^2(1-p_j)^2 \leq 4(p_k+p_j).$$

Indeed,

$$\begin{aligned}
1-(1-p_k)^2(1-p_j)^2 &= 1-(1+p_kp_j)^2-(p_j+p_k)^2+2(1+p_kp_j)(p_j+p_k)\\
&\leq 1-(1+p_kp_j)^2+2(1+p_kp_j)(p_j+p_k)\\
&= -p_kp_j(2+p_kp_j)+2(1+p_kp_j)(p_j+p_k)\\
&\leq 2(1+p_kp_j)(p_j+p_k)\\
&\leq 4(p_j+p_k)
\end{aligned}$$

Then we have in the second term:

$$4\frac{g^4(n)}{n^4}\sum_{j\neq k} COV(\mathbb{1}_{[X_k=2]}, \mathbb{1}_{[X_j=2]})$$

$$= 4\frac{g^4(n)}{n^4}\sum_{j\neq k}\Big(E(\mathbb{1}_{[X_k=2]}\mathbb{1}_{[X_j=2]}) - E(\mathbb{1}_{[X_k=2]})E(\mathbb{1}_{[X_j=2]})\Big)$$

$$= 4\frac{g^4(n)}{n^4}\sum_{j\neq k}\Big(E(\mathbb{1}_{[X_k=2],[X_j=2]}) - E(\mathbb{1}_{[X_k=2]})E(\mathbb{1}_{[X_j=2]})\Big)$$

$$= 4\frac{g^4(n)}{n^4}\sum_{j\neq k}\Big(\frac{n!}{4(n-4)!}p_j^2p_k^2(1-p_j-p_k)^{n-4} - \frac{n!}{2(n-2)!}p_j^2(1-p_j)^{n-2}$$

$$\cdot\frac{n!}{2(n-2)!}p_k^2(1-p_k)^{n-2}\Big)$$

$$= \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left(\frac{n!}{4(n-4)!}p_j^2p_k^2(1-p_j-p_k)^{n-4} - \left(\frac{n!}{2(n-2)!}\right)^2 p_j^2p_k^2(1-p_j)^{n-2}(1-p_k)^{n-2}\right)$$

$$= \frac{4g^4(n)}{n^4}\sum_{j\neq k}\Big(\frac{(n-3)(n-2)(n-1)n}{4}p_j^2p_k^2(1-p_j-p_k)^{n-4} - \left(\frac{(n-1)^2n^2}{4}\right)$$

$$.p_j^2p_k^2(1-p_j)^{n-2}(1-p_k)^{n-2}\Big)$$

$$\leq \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left(\frac{(n-1)^2n^2}{4}p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4} - \left(\frac{(n-1)^2n^2}{4}\right)p_j^2p_k^2(1-p_j)^{n-2}(1-p_k)^{n-}\right.$$

$$= \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left(\frac{(n-1)^2n^2}{4}p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}\left(1-(1-p_j)^2(1-p_k)^2\right)\right)$$

$$\leq \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left(\frac{(n-1)^2n^2}{4}p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}4(p_k+p_j)\right)$$

$$\leq \frac{4g^4(n)}{n^4}\sum_{j\neq k}\Big(n^4p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}(p_k+p_j)\Big)$$

$$\leq 4g^4(n)\sum_{j\neq k}\Big(p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}(p_k+p_j)\Big)$$

By symmetry

$$= 8g^4(n)\sum_{j,k}\Big(p_j^3p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}\Big)$$

$$= 8\left(g^2(n)\sum p_k^3(1-p_k)^{n-4}\right)\left(g^2(n)\sum p_k^2(1-p_k)^{n-4}\right)$$

We see that $g^2(n)\sum p_k^2(1-p_k)^{n-4}$ converges to 0 by condition 4.1.
But to see that the second factor converges also to 0, let $f(p)=p^2(1-p)^{n-4}$ and compute its maximum value.

$$\begin{aligned} f'(p) &= 2p(1-p)^{n-4}+(n-4)p^2(1-p)^{n-5}(-1) \\ &= p(1-p)^{n-5}\left(2+p(-n+2)\right) \end{aligned}$$

$$f'(p)=0 \iff 2+p(-n+2)=0 \iff p=\frac{2}{n-2}$$

After computing the second derivative of $f(p)$ at $\frac{2}{n-2}$, then we deduce that $f(p)$ attains its maximum at this point where $p\in(0,1)$
then:

$$\begin{aligned} f(p_k) &\leq f\left(\frac{2}{n-2}\right) \\ p_k^2(1-p_k)^{n-4} &\leq \left(\frac{2}{n-2}\right)^2\left(1-\frac{2}{n-2}\right)^{n-4} \end{aligned}$$

multiply both sides by $g^2(n)p_k$:

$$g^2(n)p_k^3(1-p_k)^{n-4} \leq g^2(n)\left(\frac{2}{n-2}\right)^2\left(1-\frac{2}{n-2}\right)^{n-4} p_k$$

and then:

$$\begin{aligned} g^2(n)\sum p_k^3(1-p_k)^{n-4} &\leq g^2(n)\left(\frac{2}{n-2}\right)^2\left(1-\frac{2}{n-2}\right)^{n-4}\sum p_k \\ &\leq \frac{4g^2(n)}{(n-2)^2} \to 0, \text{ as } n\to\infty \end{aligned}$$

Hence $V(\hat{c}_2)\to 0$ as $n\to\infty$
and $P\left(|\hat{c}_2-c_2|\geq\epsilon\right)\to 0$ as $n\to\infty$
then $\hat{c}_2\to c_2$ as $n\to\infty$ and $\hat{c}_2$ is consistent.

Now,

$$\begin{aligned}
\frac{n(C'-C)}{\sqrt{N_1+2N_2}} &= \frac{n\sqrt{c_1+c_2}}{g(n)\sqrt{N_1+2N_2}} \times \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}} \\
&= \frac{\sqrt{c_1+c_2}}{\sqrt{\frac{g^2(n)}{n^2}N_1+2\frac{g^2(n)}{n^2}N_2}} \times \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}} \\
&= \frac{\sqrt{c_1+c_2}}{\sqrt{\hat{c}_1+\hat{c}_2}} \times \frac{g(n)(C'-C)}{\sqrt{c_1+c_2}}
\end{aligned}$$

Since $\hat{c}_1 \to c_1$ and $\hat{c}_2 \to c_2$ as $n \to \infty$

then $\hat{c}_1 + \hat{c}_2 \to c_1 + c_2$

and $\frac{\sqrt{c_1+c_2}}{\sqrt{\hat{c}_1+\hat{c}_2}} \to 1$ as $n \to \infty$

Use now Slutskey's theorem to prove that:

$$\frac{n(C'-C)}{\sqrt{N_1+2N_2}} \xrightarrow{D.} N(0,1) \text{ as } n \to \infty$$

which finishes the proof.

We note that the condition of Theorem 4.2 and 4.3 requires no further knowledge of $g(n)$ other than its existence.

□

Theorem 4.3 leads to an approximate $(1-\alpha)$ level confidence interval for C:

$$\left(1-\frac{N_1}{n}\right) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}$$

where $z_{\alpha/2}$ is the usual constant for a normal confidence interval.

Proof. By theorem 4.3:

$$\frac{n(C'-C)}{\sqrt{N_1+2N_2}} \xrightarrow{D.} N(0,1)$$

For n large:

$$
\begin{aligned}
1-\alpha =& P\left(-z_{\frac{\alpha}{2}} < \frac{n(C'-C)}{\sqrt{N_1+2N_2}} < z_{\frac{\alpha}{2}}\right) \\
=& P\left(-z_{\frac{\alpha}{2}} < \frac{C'-C}{\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}} < z_{\frac{\alpha}{2}}\right) \\
=& P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < C'-C < z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right) \\
=& P\left(-C'-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < -C < -C'+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right) \\
=& P\left(C'-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < C < C'+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right) \\
=& P\left(\left(1-\frac{N_1}{n}\right)-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < C < \left(1-\frac{N_1}{n}\right)+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right)
\end{aligned}
$$

Therefore, for n large, the approximate $(1-\alpha)$ level confidence interval for C is:

$$\left(1-\frac{N_1}{n}-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}},\ 1-\frac{N_1}{n}+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right)$$

and for completeness, an approximate $(1-\alpha)$ level confidence interval for π_0 is:

$$P\left(\left(1-\frac{N_1}{n}\right)-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < 1-\pi_0 < \left(1-\frac{N_1}{n}\right)+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right) = 1-\alpha$$

$$P\left(\frac{N_1}{n}-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}} < \pi_0 < \frac{N_1}{n}+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right) = 1-\alpha$$

Therefore, the approximate $(1-\alpha)$ level confidence interval for π_0 is:

$$\left(\frac{N_1}{n}-z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}},\ \frac{N_1}{n}+z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2}+\frac{2N_2}{n^2}}\right)$$

□

Example 4.2 Use the data in Table 1 to construct a 95% confidence interval for π_0. With $n = 2000$, $N_1 = 12$ and $N_2 = 1$:

$$\frac{N_1}{n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{N_1}{n^2} + \frac{2N_2}{n^2}} = \frac{12}{2000} \pm 1.96\sqrt{\frac{12 + 2(1)}{2000^2}}$$
$$= 0.006 \pm 0.0037$$
$$= (0.0023,\ 0.0097)$$

Chapter 5

Conclusion

The sufficient condition of this thesis and Example 1 together ensure the existence of a non-degenerated asymptotic normality law for a non empty class of distributions.

A similar result was obtained by Esty [15] but it did not establish a non-degenerated normality law for a fixed $\{p_k\}$ as already explained above.
Esty [15] established a $\sqrt{n}$-normality law for $C' - C$ but allowing the underlying probability distribution to vary within a family $\left\{\{p_k\}_m : m = 1, 2, \cdot\right\}$ as the sample size n increases.

For the method of proof, we used a direct evaluation of the characteristic function of the normalized coverage probability, with an appropriate partition that allows us to control the tail probabilities.
Although the sufficient condition of Esty [15] and that of this thesis describe different populations, an intuitive comparison is still possible. Esty's condition is essentially a thicker tail condition. It says that as n increases, the total probability of unobserved species does not converge to zero but inflates at a rate such that the total probability remains constant. On the other hand, condition (4.1)

allow the total probability to converge to zero. It is therefore conceivable, in some sense, that the respective biases converge to zero at different rates, slower under Esty's condition, and faster under Condition (4.1). The difference is reflected by the fact that the rate of convergence $g(n)$ is higher than $\sqrt{n}$.

Finally, only the existence of $g(n)$ is needed, we were able to use the main results to carry out statistical inference including hypothesis testing and construction of a confidence interval for π_0.

Bibliography

[1] B. Efron and R. Thisted, Estimating the number of unseen species: how many words did Shakespeare know?, *Biometrika*,63 (1976), page 435-447.

[2] R. Thisted and B. Efron, Did Shakespeare write a newly-discovered poem?, *Biometrika*,74 (1987) page 445 - 455.

[3] I.J. Good and G.H. Toulmin, The number of new species, and the increase in population coverage, when a sample is increased, *Biometrika*,43 (1956) page 45 - 63.

[4] A. Chao,On estimating the probability of discovering a new species, *Ann. Stat.* 9 (1981) page 1339 - 1342.

[5] C.X. Mao and B.G. Lindsay, A Poisson model for the coverage problem with a genomic application,*Biometrika* 89 (2002) page 669-681.

[6] C-H. Zhang, Estimation of sums of random variables: Examples and information bounds, *Ann. Stat.* 33 (2005), page 2022 - 2041.

[7] I.J. Good, The population frequencies of species and the estimation of population parameters, *Biometrika* 40 (1953) page 237 - 264.

[8] B. Harris, Determining bounds on integrals with applications to cataloging problems, *Ann. Math. Stat.* 30 (1959) page 521 - 548.

[9] —, Statistical inference in the classical occupancy problem unbiased estimation of number of classes, *J. Am. Stat. Assoc.* 63 (1968) page 837 - 847.

[10] H.E. Robbins, Estimating the total probability of the unobserved outcomes of an experiment, *Ann. Math. Stat.* 39 (1968) page 256 - 257.

[11] N. Starr, Linear estimation of probability of discovering a new species, *Ann. Stat.* 7 (1979) page 644 - 652.

[12] L. Holst, Some asymptotic results for incomplete multinomial or Poisson samples, *Scand. J. Stat.* 8 (1981) page 243 - 246.

[13] A. Chao, Nonparametric of the number of the classes in a population, *Scand. J. Stat.* 11 (1984) page 265 - 270.

[14] W.W. Esty, Confidence intervals for the coverage of the low coverage samples, *Ann. Stat.* 10 (1982) page 190 - 196.

[15] —, A normal limit law for a nonparametric estimator of the coverage of a random sample, *Ann. Stat.* 11 (1983) page 905 - 912.

[16] —, Estimation of the number of classes in a population and the coverage of a sample, *Math. Sci.* 10 (1985) page 41 - 50.

[17] —, The size of a coinage, *Numis. Chron* 146 (1986a) page 185 - 215.

[18] —, The efficiency of Good's nonparametric coverage estimator, *Ann. Stat.* 14 (1986b) page 1257 - 1260.

[19] A. Chao and S. Lee, Estimating the number of classes via sample coverage, *J. Am. Stat. Assoc.* 87 (1992) page 210 - 217.

[20] M.S. Bartlett, The characteristic function of a conditional statistic, *J. Lond. Math. Soc.* 13 (1938) page 62 - 67.