Bayesian methods to estimate sub - population

النماذج البيزية في تقدير المجتمعات المخفية

أ. د. قتيبة نبيل نايف/كلية الادارة والاقتصاد/جامعة بغداد
الباحث/ حسين حميد خلف

2019

OPEN ACCESS

P - ISSN 2518 - 5764
E - ISSN 2227 - 703X

Received:21/5/2018
Accepted: 24/6/2018

المستخلص

يهدف البحث الى تقدير المجتمعات السكانية المخفية اذ تم تقدير اعداد متعاطي الحبوب المخدرة في بغداد للذكور للفئة العمرية (15- 60) ، بالاعتماد على نماذج بيز والتي من خلالها يتم معالجة بعض التحيز التي تعاني منها الطريقة الشائعة الاستعمال (*Killworth method*) في التقارير المعتمدة في العديد من دول العالم .

تم استعمال على 4 نماذج وهي (الدرجة العشوائية (*Random degree*) ، تأثيرات الميول (*Barrier effects*)، الافصاح المتحيز(*Transmission bias*) ، المشترك (*Combined*)) ، يعد انموذج الدرجة العشوائي بمثابة امتداد لأنموذج *Killworth* عند اضافة التأثيرات العشوائية كالتباين وعدم اليقين عبر حجم الشبكة الشخصية، وعند توسيعه باضافة احتساب ان للمستجيبين ميول مختلف فيما بينهم ، ان هذه الاضافة تتم بخلط المتغيرات غير العشوائية مع العشوائية لينتج أنموذج تأثيرات الميول ، اما الامتداد الاخر لأنموذج الدرجة العشوائية يكون باضافة قلة وعي المستجيبين لينتج أنموذج الافصاح المتحيز، ولغرض تحسينه يكون باضافة معلومات اخرى ، واخيراً الأنموذج المركب او المشترك الذي يجمع بين الأنموجين الثاني والثالث .

تبين ان انموذج الدرجة العشوائية هو افضل النماذج وفق البيانات التي جمعت باستعمال استبانة احصائية اعدت لهذا الغرض فقد لوحظ انه الافضل بالاعتماد على مؤشر الكفاءه النسبية، وان نتائج المحاكاة كانت داعمة لهذا الاستنتاج ، ولقد بلغ تقدير اعداد المتعاطين وفق هذا الاسلوب 32,862 شخصاً من الذكور في الفئة العمرية (15 – 60) .

المصطلحات الرئيسة للبحث / NSUM ، السكان المختبئون ، (طرق المتابعة ، أساليب كيلورث ، الحاجز وناقل الحركة).

Journal of Economics and Administrative Sciences
2019; Vol. 25, No.110
Pages: 406- 425

*البحث مستل من رسالة ماجستير

المقدمة (Introduction)

تعاني العديد من البلدان الغنية منها والفقيرة على حد سواء من وجود مجاميع سكانية متداخلة ضمن مجتمعاتها يصعب تقديرها مثل اعداد متعاطي المخدرات او اعداد المصابين بامراض نقص المناعة المكتسبة (الايدز) لعدة اسباب اهمها ان العديد من هذه المجاميع تكون مغلقة على نفسها لا المنتمي اليها لا يفصح عن انتماءه خوفا من الملاحقات القانونية والوصم المجتمعي (Stigma) له ، كذلك يميل اغلب الناس الى تكوين علاقات مع من يشابهونهم في الصفات والطباع ، ان هذه الاسباب وضعت الصعوبات امام المسوحات التقليدية لذا تم اللجوء الى اساليب اخرى في عملية التقدير ، وكان $(\boldsymbol{Bernard \& et\ al})$ عام 1989 م هو اول من شرع بتقدير مجموعة صعبة التقدير وهم ضحايا الزلزال الذي ضرب المكسيك [I] عام 1985 م ، اذ ادرك ان كل فرد يمتلك معلومات عن شبكته الشخصية وباستطاعته التقدير من خلال احتساب حجم الشبكة الشخصية $(\boldsymbol{personal\ network\ size})$ والتي هي اعداد الاشخاص الذين يعرفهم المستجيب كالاصدقاء والاقارب والمحددة مكانياً وزمنياً حسب طبيعة الدراسة [V] ، ان هذا الاسلوب هو البداية الحقيقة لتقدير المجتمعات المخفية ولاحتساب حجم شبكة شخص ما عن طريق سؤال الشخص المعني ان كان يعرف بعض الاشخاص من المجتمع مسحوبين بطريقة عشوائية وعليه سوف تخضع هذه الالية الى توزيع ثنائي الحدين $(\boldsymbol{Binomial\ distribution})$ وبعدد n من المحاولات وعلية سوف يكون الاحتمال كالاتي : $\boldsymbol{d_i/_N}$ ، ويمكن زيادة دقة التقدير وذلك بأخذ معدل مجموعة مستجيبين والاستفسار عن مجاميع جزئية مختلفة وهذا هو مقدر $(\boldsymbol{Killworth})$ $\boldsymbol{scale-up\ estimator}$.

ان النماذج الاخرى سوف تبنى على أنموذج $(\boldsymbol{Killworth})$ وكالاتي أنموذج الدرجة العشوائي $(\boldsymbol{Random\ degree})$ وهو امتداد لأنموذج $(\boldsymbol{Killworth})$ باضافة التأثيرات عشوائية كالتباين وعدم اليقين عبر حجم الشبكة الشخصية، وعند توسيعه باضافة احتساب حقيقة ان المستجيبين لديهم ميول مختلفة فيكون بخلط المتغيرات غيرالعشوائية مع العشوائية لينتج أنموذج تأثيرات الميول $(\boldsymbol{Barrier\ effects})$، اما الامتداد الاخر لأنموذج الدرجة يكون باضافة قلة وعي المستجيبين لينتج أنموذج الافصاح المتحيز $(\boldsymbol{Transmission\ bias})$، ولغرض تحسينه يكون باضافة معلومات اخرى، واخيراً الأنموذج المركب او المشترك $(\boldsymbol{Combined})$ الذي يجمع بين الأنموجين الثاني والثالث. [VIII]

$\underline{\boldsymbol{Network\ Scale-Up\ Method\ (NSUM)}}$

هي الية مصممة لتقدير حجم المجاميع السكانية المخفية والصعبة الوصول بدون الحاجة الى اجراء مقابلات مباشرة مع افراد تلك المجتمعات انما يكون عن طريق عينة ممثلة للسكان والذي يكون المجتمع المخفي جزء منه ، ومن هذه العينة يمكن الاستفسار عن عدد الاشخاص الذين ينتمون للمجموعة المستهدفة ضمن شبكة معرفة المستجيبين ، ويمتاز هذا الاسلوب بـما يلي :

1- لا يَسأل عن الصفات الشخصية للمستجيب والانتماء لمجموعة وصم ام لا.

2- يمكن استعمال اسهل اساليب المعاينة الشائعة واقلها كلفة في المسوحات . [IV]

الافتراضات التي تستند عليها $\underline{(\boldsymbol{NSUM})}$ وكالاتي : -

1- ان جميع افراد المجتمع لهم فرص متساوية لمعرفة الاشخاص في المجموعات الجزئية وهذه الفرصة تزيد عندما يزداد حجم المجتمع المستهدف ولغرض معالجة عدم تحقق الشرط سيتم بناء أنموذج تأثيرات الميول .

2- ان كل مستجيب له معرفة جيدة عن جميع الناس الذين يعرفهم (كأن يعرف جميع الاشخاص ضمن شبكة معرفته والذين هم يتعاطون شيئاً مخدراً)، وهنا يكون أنموذج الأفصاح المتحيز هو الضامن لحل هذه المشكلة .

3-المستجيب يمكنه ان يجيب بدقة وبوقت قصير عن الافراد الذين ينتمون لمجموعه ما ، ولديه اتصال معهم بأحدى الوسائل التي ذكرت . [IX]

ان عدم تحقق احدى هذه الشروط سوف يجعلنا نبني أنموذجا اخر حسب الشرط الذي انتهك، وان عملية البناء تعتمد على تطوير اطار بيز $(Bayesian\ Framework)$ لتقدير عدد السكان بأسلوب $NSUM$().

مشكلة البحث

ان انموذج $(Killworth)$ الشائع الاستعمال في تقدير المجتمعات المخفية الا ان هذا الانموذج يعاني من تأثير بعض انواع التحيزات وهما اولا تأثير الميول حيث يميل المستجيب الى تكوين علاقات مع اشخاص يشابهونهم بالصفات اما التأثير الاخر فهو الافصاح حيث ان المستجيبين لا يصرحون بتعاطيهم الحبوب المخدرة .

الهدف

- تقدير اعداد اشخاص واحداً من المجتمعات السكانية المخفية والصعبة الوصول اليه وهم متعاطو الحبوب المخدرة بعدة طرائق بيزية .
- بناء نماذج لتقدير المجتمعات السكانية والمقارنة بينها وأيهما اكثر ملائمة للتقدير في المجتمع العراقي

النماذج المستعملة للتقدير:

أ - أنموذج $(Killworth)$

جاءت تسميته نسبةً الى العالم بهذا التخصص وهو $(Peter\ Killworth)$ ، ويعد هذا الانموذج هو الاكثر استعمالا في التقدير، ويتكون هذا الانموذج نتيجة الدمج بين تقدير الشبكة الشخصية واجابة المستجيبين عن السؤال الخاص بالمجتمعات المخفية، اذ يبدأ الانموذج بتقدير حجم الشبكة الشخصية وفق المعادلة رقم (1) وكالاتي :

$$\hat{d}_i = \frac{\sum_{k=1}^{K-1} y_{ik}}{\sum_{k=1}^{K-1} N_k} * N \qquad (1)$$

اذ ان :

y_{ik} : عدد الاشخاص المعروفين من قبل المستجيب i في المجموعة الجزئية k

N_k : حجم المجموعة السكانية k والتي تاخذ من المسوحات والتقارير الحكومية

N : حجم السكان الكلي يأخذ من نتائج التعداد [X]

اذ يتم اخذ مجموع جميع استجابات المستجيبين $i = 1, 2, \dots, n$ عند الاستفسار عن عدد الاشخاص الذين ينتمون للمجموعات المعلومة العدد والتي يبلغ عددها $K - 1$ والتي نرمز لها y_{ik} مقسومة على مجموع المجموعات المعلومة العدد والتي يتم الحصول عليها من الاجهزة الحكومية والمسوحات الدورية ثم يضرب الناتج في اجمالي عدد السكان، يكون الناتج النهائي متجه من حجم الشبكات الشخصية ذو بعد $(n * 1)$.

بعد ذلك يتم تقدير عدد الاشخاص الذين ينتمون للمجموعة المخفية والتي تعتبر هي المجموعة الاخيرة ضمن قاعدة البيانات والتي يرمز لها K ، وبما ان $i\hat{d}$ هي تقديرات شرطية للـ d_i فيكون مقدر دالة الامكان الاعظم الى N_K (حجم السكان غير المعروف) ويحسب وفق المعادلة الأتية :

$$\hat{N}_K = N * \frac{\sum_{i=1}^{n} y_{iK}}{\sum_{i=1}^{n} \hat{d}_i} \qquad (2)$$

اذ يمثل البسط هنا المجموع الكلي لاعداد الاشخاص في المجموعة المخفية فقط مضروبا باجمالي عدد السكان، بينما يشير المقام الى اجمالي حجم الشبكات الشخصية للمستجيبين عندئذ تمثل القيمة النهائية لاجمالي اعداد المجموعة المخفية وفق اسلوب $(Killworth)$.

ب- النماذج البيزية : هي نماذج تم تصميمها لغرض تقدير المجتمعات السكانية المخفية بالاعتماد على توزيعات مفترضة وتوزيعات اولية تم احتسابها بالاعتماد على معلومات سابقة، تم استعمال 4 نماذج منها لمعالجة حالات معينة من التحيزات وهي :-

اولا- أنموذج الدرجة العشوائية (Random degree model) :-

يعد هو بمثابة امتداد لأنموذج $(Killworth)$ وفيه يكون توزيع التأثير العشوائي لحجم الشبكة الشخصية (d_i) يعتمد على توزيع المشاهدات في أنموذج $(Killworth)$ المستخدمة في تقدير $\hat{d}$ ، فلو افترضنا انها تتوزع توزيع اللوغاريتمي الطبيعي $log\ normal\ dist.$ فان الأنموذج الناتج سيكون فيه :

$$y_{ik} \sim Binom(d_i , {N_k}/{N})$$

$$d_i \sim Log\ Normal\ (\mu , \sigma^2)$$

ان ألية تقدير معلمات الأنموذج تكون وفق أسلوب بيز $(Bayesian\ procedure)$ بالاعتماد على التوزيعات السابقة $(prior\ distributions)$ لكل من المتوسط والانحراف المعياري

$$\mu \sim U\ (3\ ,8\ )$$

$$\sigma \sim U\ (\frac{1}{4}\ ,2\ )$$

ان القيم الاولية $(priors)$ الى $(\mu,\ \sigma)$ تم التوصل لهما من العديد من البيانات ومن دول متعددة منها (الولايات المتحدة، اوكرانيا، مولدافيا، كازخستان، البرازيل) وتكون وفق التوزيع المنتظم $(uniform)$ وتعد قيم قياسية [VIII] ، فلو اردنا ان نغير هذا المفهموم فيجب ان نفترض قيمة ابتدائية تكون خارج التوزيع المنتظم ثم تعامل بأسلوب سلسلة ماركوف مونتي كارلو $(MCMC)$ $(Markov\ Chain\ Monte\ Carlo)$ لغاية الحصول على المدى المطلوب .

وعليه يكون التوزيع اللاحق المشترك المشروط $(Joint\ Conditional\ posterior\ distribution)$ لهذا الأنموذج وفقا الافتراضات السابقة موضحة بالمعادلة ادناه

$$\pi(\mu,\ \sigma^2, d_i , N_K\ \mid\ y_{ik} , N_k\ , N) \propto$$
$$\left(\prod_{i=1}^{n} \left[\frac{1}{d_i\ \sigma\sqrt{2\pi}}\ e^{-\frac{(\log(d_i)-\mu)^2}{2\sigma^2}}\ \prod_{k=1}^{K} \left(\binom{d_i}{y_{ik}} \left(\frac{N_k}{N}\right)^{y_{ik}} \left(1-\frac{N_k}{N}\right)^{d_i-y_{ik}} \right) \right] * \frac{1}{N_K} * \frac{1}{5} * \frac{1}{\frac{7}{4}} \right. \ldots\ldots (3)$$

ولغرض اجراء تحديث المعلمات سيتم تحديث كل معلمة بدلالة الاخرى .

أ : تحديث قيمتي الوسط الحسابي والتباين

لحساب قيمة التباين σ^2 نستعمل معاينة جبس $(Gibbs\ sampler)$ فيكون التوزيع اللاحق الشرطي متقارب جدا من توزيع معكوس كاما $(inverse\ gamma)$ ، بينما يكون التوزيع السابق محدد مسبقا عن طريق معلومات موجودة من بيانات مسوحات او دراسات سابقة ، وعليه يكون التوزيع اللاحق الى تباين المجتمع σ^2 كالاتي :

$$\pi(\sigma^2 \mid \mu,\ N_K , d_i , y_{ik} , N_k\ , N)$$
$$\sim Inverse\ Gamma\ \left(\frac{n-1}{2} , \frac{\sum_{i=1}^{n}(\log(d_i) - \mu)^2}{2} \right) I_{(\frac{1}{4}<\sigma<2)} \ldots\ldots (4)$$

اما التوزيع اللاحق المشروط للمتوسط μ فيصبح توزيع طبيعي مبتور $(truncated\ normal)$(يتم احتسابه بنفس الاسلوب السابق باستعمال معاينة جبس $(Gibbs\ sampler)$ لتحديث μ وكالاتي :

$$\pi(\mu \mid \sigma^2 , N_K, d_i , y_{ik} , N_k , N) \sim Normal\left(\frac{\sum_{i=1}^{n} \log d_i}{n} , \frac{\sigma^2}{n}\right) I_{(3<\mu<8)} \ldots .(5)$$

ان القيم المتولدة للوسط الحسابي والانحراف المعياري يجب ان تقع ضمن مدى القيم الاولية .
ان القيمة الاولية الى μ خصصت لمتوسط عدد متغيرات تصل الى 3000 متغير ضمن مدى مجموعة البيانات وهو ما يتفق مع بحوث سابقة اجريت في مجال الشبكات الاجتماعية و $(NSUM)$.

اما القيمة الاولية الى σ فانها تسمح لـ 95% من قيم المتولدة ان تكون ضمن المدى المطلوب عند مضاعفات الانحراف المعياري بقيم تتراوح (من 1.6 الى 55) من المرات في كلا الاتجاهين للمتوسط والتي تغطي بالكامل نطاق النتائج من تقديرات $(killworth)$ عبر مجموعة بيانات متعددة [VIII].
اما القيمة المستخدمة كمضروب ثابت (معلمة ضبط) عند التوليد المفترض هي 2.3 والمقترحة من قبل $(Gelman)$ واخرون عام 1995 والتي تعد كقيمة مثلى عند التعامل بها لكي يكون التباين في أمثل حالة عندما نحاكي التوزيع الطبيعي القياسي ذو متغير واحد باستخدام خوارزمية ميتروبلس $(Metropolis)$ ، ان آلية تحديد هذه القيمة تمت من خلال النظر الى معيارين مختلفين من المعايير المثلى و هما : [XI]
1- كفاءة التقارب $(asymptotic\ efficiency)$ لسلسلة ماركوف .
2- معدل التقارب الهندسي $(geometric\ convergence\ rate)$ لسلسلة ماركوف .
ولقد تم اثبات فعالية هذه القيمة من قبل $(Raftery\ \&\ Lewis)$ [XI] في مصفوفة التباين المشترك اللاحقة والشائعة في معلمات الانحدار اذ تبين ان المصفوفة نتيجة المحاكاة تبقى متقاربة تلك الموجودة في العينة .
ان هذه القيمة سوف تستعمل كمعلمة ضبط $(tuning\ parameter)$ في توليد جميع المعلمات الاخرى بواسطة التوزيع الطبيعي ذي معلمتين مقيدتين . في جميع النماذج يتم تقدير المعلمات بالاعتماد على طريقة سلسلة ماركوف مونتي كارلو $(MCMC)$ الا ان هاتين المعلمتين تقدران بواسطة معاينة جبس $Gibbs$ بينما المعلمات الاخرى عن طريق خطوات ميتروبولس العشوائية $(Random\ Walk\ Metropolis)$ ، وهنا يعتمد الاعتماد على أنموذج $(Killworth)$ كنقاط شروع لهذه المعلمات. ان خوارزميات سلسلة ماركوف مونتي كارلو $(MCMC)$(تعتمد على سلسلة شروع اولية لتقدير الانحراف المعياري للتوزيع الشرطي اللاحق لكل معلمة ، وتستخدم معلمة الضبط للانحراف المعياري عند الشروع بأليه التوليد عن طريق التوزيع الطبيعي [VIII] ، سيتم اعتماد على هاتين القيمتين في جميع النماذج .
وان معلمة الضبط $(tuning\ parameter)$: هي قيمة مستخدمة من الخوارزمية الاحصائية من اجل السيطرة على سلوكها ، ويجب اختيارها بدقة من اجل الحصول على مقدرات ذات حصانة قوية و كفوءة[XV]
معاينة جبس $(Gibbs\ Sampler)$: هو اسلوب لتوليد سلسلة من العينات العشوائية من هامش التوزيع $(marginal\ distribution)$ بطريقة غير مباشرة أي بدون الحاجة لحساب دالة الكثافة $(the\ density\ function)$، وهو يعتمد فقط على الخصائص الاولية لسلاسل ماركوف $(Markov\ chains)$ ومن خلاله يمكن تجنب الحسابات المعقدة وابدالها بمتسلسلات من الحسابات السهلة وله تأثير واسع في حل المشاكل العملية، وأكثر استعمالاته في نماذج بيز $(Bayesian\ models)$ لغرض توليد التوزيعات اللاحقة $(posterior)$ بينما في النماذج التقليدية يستعمل لحساب دالة الامكان الاعظم $(likelihood)$ وخصائص المقدرات وله تأثير ايضا في المجال النظري . [III]

ب- تحديد حجم الشبكة الشخصية :

التوزيع اللاحق الشرطي لتقدير حجم الشبكة الشخصية (d_i) فتكون كالاتي :

$$\ell(d_i \mid \mu, \sigma^2, N_K, y_{ik}, N_k, N)$$
$$\propto -\log d_i - \frac{(\log(d_i)-\mu)^2}{2\sigma^2} + \sum_{k=1}^{K} \log\binom{d_i}{y_{ik}} + \sum_{k=1}^{K} d_i \log\left(1-\frac{N_k}{N}\right) \ldots (6)$$

يلاحظ في المعادلة اعلاه انها لا تقترب من شكل معروف من احدى التوزيعات المعروفة ، لذلك سوف نستعمل اسلوب ميتروبلس ($Metropolis\ step$) لايجاد توليد وتحديث قيم (d_i) والموضحة في الخوارزمية الاتية :

I. نولد قيم بالاعتماد على اسلوب (Killworth) على ان ترفض القيمة التي تكون ادنى من ($k^{max}\ y_{ik}$) .

II. تحذف اول العينات المتولدة بسبب ان طريق تلك القيم يكون بطيء وتدريجي وهذا يؤدي الى ارتباط ذاتي عالي للمتتالية المتولدة هذا السبب يجعلنا بحاجة الى عدد كبير من التكرارات لذلك يفضل اعتماد بيانات العينة كنقاط شروع حقيقية[XI] .

III. نولد قيم لاحقة تتبع التوزيع الطبيعي بمعلمات محددة بوسط حسابي معتمدا على التوليد السابق وبتباين يكون متأثرا بمعلمة الضبط (من خلال ايجاد الانحراف المعياري للبواقي في الانموذج) على سلسلة شروع ابتدائية (والتي بدورها تسمح بتقدير قيم انطلاق معقولة لكثير من المعلمات في طريقة سلسلة ماركوف مونتي كارلو ($MCMC$) ، من الضروري العناية عند اختيار القيم الاولية للانطلاق بالتقدير فان القيم البعيدة عن الواقع سوف تؤدي بحاجتنا الى عدد كبير من التكرارات

IV. اذا كانت القيم اللاحقة أصغر من ($\max(dat[i,]$ اذ تمثل ($dat[i,j]$) مصفوفة الاستجابات فانه سيتم الاعتماد على القيم السابقة المتولدة وفق الخطوة الاولى ، عدا ذلك سيتم تعويض القيم السابقة في المعادلة رقم (6) تارة و تعويض القيم اللاحقة تارةً اخرى .

V. نجد الفرق بين التعويضين ، نأخذ هذا الفرق لغرض المقارنة اذا كانت قيمته سالبة، في حين يأخذ الصفر اذا كانت قيمة الفرق موجبة .

VI. نولد قيمة تتبع التوزيع المنتظم ($uniform$) ضمن الفترة (1 , 0) ثم نجد اللوغاريتم الطبيعي لها لضمان ان تحمل القيمة المتولدة اشاره سالبة، ان افضل نتيجة يمكن الحصول عليها للتوزيع هو ان تكون هوامشه ($marginals$) تقترب من الطبيعي ولغرض تحسين ذلك اجراء نوع من التحويلات ($transformations$) مثل اخذ اللوغاريتمات لكل القيم الموجبة و اللوغاريتم الطبيعي للقيم المحصورة بين (0 , 1) .[III]

VII. نجري مقارنة بين القيمتين الناتجتين وفق الخطوتين (V&VI)، فاذا كانت القيمة الناتجة في الخطوة V اصغر من تلك الناتجة وفق الخطوة VI فانه سيعتمد على القيم اللاحقة عدا ذلك يتم الاعتماد على القيم المتولدة سابقاً ، وتكون النتيجة النهائية متجة ذي بعد ($n * 1$) حيث n يمثل عدد المستجيبين .

ج : تقدير حجم المجتمع المخفي N_K:

يلاحظ ان التوزيع الشرطي اللاحق لحجم المجتمع الجزئي N_K معادلة رقم (7) لا يمتلك صيغة قريبة من توزيع معروف مثلما هو الحال في تحديد حجم الشبكة الشخصية لذلك نستخدم ميتروبلس $Metropolis\ step$ لغرض توليد وتحديث المعلمات

$$\ell(N_K \mid \mu,\ \sigma^2, d_i,\ y_{ik}, N_k, N) \propto \sum_{i=1}^{n} y_{iK} \log\left(\frac{N_k}{N-N_K}\right) + \sum_{i=1}^{n} d_i \log\left(1-\frac{N_k}{N}\right) - \log N_K \quad \dots(7)$$

ان آلية التوليد تكون مشابه نوعا ما لأسلوب تحديد حجم الشبكة الشخصية للمستجيب باختلاف الشرط الموضوع لقبول القيمة المتولدة ، فالمنطق هو من يفرض نفسه كشرطاً للتحديد فليس من المنطقي ان يكون حجم المجتمع المخفي اكبر من عدد السكان الكلي او اصغر من عدد الاشخاص المتعاطين الذين يعرفهم مستجيب ما اي بمعنى يجب ان تقع القيمة المتولدة ضمن حدود الفترة $((k^{max\ y_{iK}}), N)$ ونادرا ما تخرج القيمة المتولدة عن هذه الفترة ، ولا تختلف خوارزمية التقدير والتحديث عن تلك التي سبقتها الا في الفرق المستخدم فانه يأتي نتيجة التعويض في معادلة التوزيع اللاحق رقم (7) الى N_K .

ثانيا- أنموذج تأثيرات الميول $(Barrier\ effects\ model)$:-

يسمي الأنموذج بتأثيرات الميول كونه يأتي من ميول المستجيبين للتعرف على الاشخاص بالاعتماد على خصائص معينة ، ويحدث عندما يعرف بعض المستجيبين عدد أكثر أو أقل من المتوقع ضمن شبكة معارفه والذين يصنفون ضمن مجموعة معينة، على الاغلب يميل الافراد لمعرفة الاشخاص الذين يشابهونهم في بعض الصفات كالعمر والجنس والتحصيل الدراسي وغيرها ، فمثلا ان الأمراة ذات ثلاثون عاما تعرف اعداد من النساء اللواتي انجبن اخر اثنا عشر شهرا أكثر من المتوقع وهذا ما يسمى $(overestimation)$ ، بينما تقل هذه الاحتمالية في حالة كان المستجيب رجلا يبلغ من العمر ثمانون عاما والذي يطلق عليه تسميه $(underestimation)$. [X]

وهنا سنلاحظ تشتت أعلى $(overdispersion)$ في توزيع عدد الاشخاص في المجموعة قيد الدراسة والذي سبق وذكر ان يتوزع ثنائي الحدين في $(Killworth)$ وأنموذج الدرجة العشوائي ، ان احتمالية معرفة المستجيب لاي شخص في المجموعة k يكون ثابتا ومساويا الى $\frac{N_k}{N}$ سيرمز لها ب q_{ik} ، اما هنا فيسكون لهذا الجزء توزيع وهو توزيع بيتا بسبب اختلاف عشوائية الاستجابة وبذلك يصبح الأنموذج متكون مما يأتي : [VIII]

$$y_{ik} \sim Binom(d_i, q_{ik}),$$

$$d_i \sim log\,normal\,(\mu, \sigma^2),$$

$$q_{ik} \sim Beta\,(m_k, p_k)$$

ان التوزيع اللاحق المشترك المشروط $(Joint\ Conditional\ posterior\ distribution)$ لهذا الأنموذج وفقا الافتراضات السابقة يكون كالاتي :

$$\pi(d_i, \mu,\ \sigma,\ m_K,\ \rho_k \mid\ y_{ik}, N, m_k) \propto$$

$$\left(\prod_{i=1}^{n}\left[\frac{1}{d_i\ \sigma\sqrt{2\pi}}\, e^{-\frac{(\log(d_i)-\mu)^2}{2\sigma^2}} \prod_{k=1}^{K}\binom{d_i}{y_{ik}} \frac{B\left(m_k\left(\frac{1}{\rho_k}-1\right)+y_{ik}\,،\, d_i+(1-m_k)\left(\frac{1}{\rho_k}-1\right)-y_{ik}\right)}{B\left(m_k\left(\frac{1}{\rho_k}-1\right)\,،(1-m_k)\left(\frac{1}{\rho_k}-1\right)\right)}\right] \frac{1}{m_K}\,\frac{1}{1}\,\frac{1}{5}\,\frac{1}{7/4} \quad \dots($$

وباستخدام توزيع بيتا - باينوميل $(beta - binomial)$ ودمج q_{ik} بفعالية وتقليل عدد المعلمات لاخذ عينات ، ان اسلوب $(MCMC)$ يجعل تحديث كل من $(\mu\ ،\ \sigma)$ مشابه للتوليد في أنموذج الدرجة العشوائي.

يلاحظ من المعادلة رقم (9) الخاصة بالتوزيع اللاحق المشروط الى m_K عدم وجود صيغة قريبة منه لغرض التحديث لذلك يستعمل خطوات ميتروبلس ($Metropolis$) كأسلوب لتحديث المعلمات مماثلا تلك التي تم تحديثها في انموذج الدرجة العشوائية .

$$\ell\left(m_K \mid y_{ik}, N, d_i, \sigma^2, \mu, m_k, \rho_k\right) \propto \sum_{i=1}^{n} \log B\left(m_K\left(\frac{1}{\rho_K}-1\right)+y_{iK}, d_i+(1-m_K)\left(\frac{1}{\rho_K}-1\right)-y_{iK}\right)-\sum_{i=1}^{n} \log B\left(m_K\left(\frac{1}{\rho_K}-1\right),(1-m_K)\left(\frac{1}{\rho_K}-1\right)\right)-\log(m_K) \quad \dots(9)$$

ان حدود m_K هي (0 , 1) والتي تمثل نسبة المجموعة الجزئية الصعبة الوصول K الى المجموع الكلي للسكان ، وهنا سنعتمد انعكاس التماثل الطبيعي ($normal\ symmetric\ reflecive$) لتلك الحدود خاصة اذا كانت القيم المتولدة خارج تلك الحدود مثال ذلك لو كانت قيمة $m_K^{(t)} = 0.9$ وكانت القيمة اللاحقة هي $m_K^{(t+1)} = 1.05$ ، فلو اضفنا 0.1 للحدود فانها بالتالي تصبح مساوياً الى 1 الا ان القيمة تبقى خارج الحدود لذلك نعمد الى الارتداد بقيمة 0.1 لتصبح ($m_K^{(t+1)} = 0.95$) لان التوزيع متماثلا ، وان تحديث ρ_k مقارب جدا لما سبقه الا ان الفرق الوحيد هو القيمة الاولية ($prior$) ، اما التوزيع اللاحق المشروط الى ρ_k هو :

$$\ell(\rho_k \mid y_{ik}, N, d_i, \sigma^2, \mu, m_k, m_K) \propto \sum_{i=1}^{n} \log B\left(m_K\left(\frac{1}{\rho_K}-1\right)+y_{iK}, d_i+(1-m_K)\left(\frac{1}{\rho_K}-1\right)-y_{iK}\right) - \sum_{i=1}^{n} \log B\left(m_K\left(\frac{1}{\rho_K}-1\right),(1-m_K)\left(\frac{1}{\rho_K}-1\right)\right) \quad \dots\dots(10)$$

ان حدود ρ_k هي مشابه لحدود m_K والتي تكون بين (0 , 1) وبنفس ما ذكر سابقا يستخدم انعكاس التماثل الطبيعي مع استخدم 2.3 كمعلمة الضبط .

ولغرض تحديث d_i يمكن تبسيط دوال بيتا و حيث انها تظهر مرة واحدة في أنموذج بيتا ، وبذلك يصبح التوزيع اللاحق له كالاتي:

$$\ell(d_i \mid y_{ik}, N, \sigma^2, \mu, m_k, m_K, \rho_k) \propto -\log(d_i) - \frac{(\log(d_i)-\mu)^2}{2\sigma^2} + \sum_{k=1}^{K} \log\left(\frac{d_i}{y_{ik}}\right) + \log\Gamma\left(d_i+(1-m_k)\left(\frac{1}{\rho_k}-1\right)-y_{ik}\right) - \log\Gamma\left(d_i+\left(\frac{1}{\rho_k}-1\right)\right) \quad \dots\dots\dots(11)$$

كما في أنموذج الدرجة العشوائية يجب ان تكون قيمة d_i اكبر من ($k^{max\ y_{ik}}$) ، وايضا يتم استخدام الفرض الاعتيادي مع معلمة الضبط المساوة لـ 2.3 من المرات لخطأ الباقي القياسي (residual standard error) يتم الحصول من خلال الانحدار (regressing) على سلسلة البداية الاولية . اما فيما يخص المعلومات الاولية (priors) الى μ، σ تبقى مشابه لتلك الموجودة في أنموذج الدرجة العشوائي [VIII]

ان المقدار q_{ik} يتوزع وفق توزيع بيتا بالمعلمتين (m_k, p_k) غير القياسية ($nonstandard\ parameterization$) لتوزيع بيتا فاذا كانت دالة الكثافة الاحتمالية لها كالاتي :

$$f_x(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

عندها ستكون $m = \frac{\alpha}{\alpha+\beta}$ و $\rho = \frac{1}{1+\alpha+\beta}$ وعليه سيكون m_k هو القيمة الاولية للمتوسط ($the\ prior\ mean$) لقيمة q_{ik} بينما يكون p_k القيمة الاولية للتشتت .

أ : تحديث قيمة m_K :

ان m_K تمثل التوقع الى (q_{ik}) أي ان :

$$E\,(q_{ik}) = m_K = {}^{N_K}/_{N} \quad \ldots..(12)$$

وهي تقع ضمن الفترة بين (1،0) كونها مساويه للجزء مقسوما على الكل ويتم تحديثها وفق الخوارزمية الاتية :

I. القيم الاولية تكون ناتجة وفق المعادلة رقم (12)
II. نولد قيم لاحقة تتبع التوزيع الطبيعي بوسط حسابي يعتمد على القيم الاولية وبتباين مضروب بمعلمة الضبط البالغة 2.3
III. يستعمل اسلوب انعكاس التماثل الطبيعي
IV. يتم تعويض القيم السابقة في المعادلة رقم (9) تارة و تعويض القيم اللاحقة تارةً اخرى .
V. نولد قيمة تتبع التوزيع المنتظم $uniform$ ضمن الفترة (1 , 0) ثم نجد اللوغاريتم الطبيعي لها .
VI. نجري مقارنة بين القيمتين الناتجتين وفق الخطوتين (V&IV) ، فاذا كانت القيمة الناتجة في الخطوة IV اصغر من تلك الناتجة وفق الخطوة V فانه سيعتمد على القيم اللاحقة عدا ذلك يتم الاعتماد على القيم المتولدة سابقاً

ب: تحديث (ρ_k) :

لا يختلف تحديث (ρ_k) كثيرا عن سابقه فيختلف عنه بتعويض في المعادلة رقم (10) بدلا من رقم (9) ، فهو يماثله في اليه توليد بالاعتماد على التوزيع الطبيعي ذي معلمات قياسية ، فتقبل القيمة المتولدة اذا وقعت بين القيم الاولية والتي تأخذ التوزيع الاتي :

$\rho_k \sim U(0\,,1)$

ان من اول الافتراضات التي اقرت في حالة عدم توفر معلومة عن التوزيع الاولي $(Noninformative$ $Bayesin\ Priors)$ ولقد افترض $(Bayes/Laplace)$ عندما لا نمتلك اي معلومة حول المعلمة فيمكن استخدام التوزيع المنتظم $(Uniform$ $dis.)$ كتوزيع اولي كونه يسمح لجميع النتائج الممكنة ليكون لديها نفس الاحتمال .[XII]

في الاسلوب البيزي يتم الاعتماد على التوزيع السابق المعلوم كأساس للوصول الى التوزيع اللاحق ، وان دالة الكثافة الاحتمالية له تعتمد على معلومات سابقة توفرت نتيجة بحوث او مسوحات سابقة ، الا انه في حالة عدم توفر مثل هكذا معلومات سوف نكون بحاجة الى توزيع سابق له تأثير ضئيل على الاستدلال وهذا ما يسمى $(prior\ noninformative)$ ، اما اذا اردنا نعرفها فهي قيمة اولية تزودنا بمعلومات قليلة لها علاقة بنفس الدراسة ولها تأثير ضئيل على البيانات .[XII]

ثالثا: أنموذج الافصاح المتحيز$(Transmission\ bias\ model)$:

يسمى الانموذج بالافصاح المتحيز لانه يحدث عندما لا يرغب المبحوث او يكون متردد بالافصاح للمستجيب بأن له سلوك او شيئا مخالفا للقانون (مثل ان يكون المبحوث يتعاطى شيئا مخدرا او بحيازته سلاح غير مرخص) لعدة اسباب يعتقدها المبحوث انها تسبب له مشكلة من ناحية الوضع الامني او ان هذه المجموعة تعتبر وصمه من الجانب الديني أو الاجتماعي وبذلك سوف يقل معرفة المستجيبين للاشخاص ضمن شبكة معرفته ، ان هذا التحيز او الخطأ يحدث عندما يكون فرداً ما ضمن شبكه المعرفة لاحد المستجيبين الا انه لا يعرف ان كان هذا الفرد ينتمي لمجموعة محددة أم لا ، فعلى سبيل المثال لو ان أحدى النساء ضمن معارف المستجيب الا انه يجهل هل ان تلك الامراة قد انجبت مولوداً خلال أخر 12 شهراً أم لا ، ان الخطأ متفاوت من مجموعة الى اخرى بالاعتماد على حساسية ووضوح معلومات تلك المجموعة حيث من الصعب تحديد هذا الخطأ .[X]

ويعزى اسباب تحيز الافصاح الى ما يأتي :

- قلة المعرفة حيث ان هذه المعلومة لم تذكر ضمن احاديث المستجيب مع المبحوث (مثلا ان كان للمبحوث شقيق توأم ام لا) .
- الوصم أو الاحراج (مثل مجموعه متعاطي المخدرات)
- المعلومات الشخصية والتي عادة لا يصرحها بها المبحوث لعديد من الناس (مثل الدخل والوزن)[V]

اسلوب احتساب قيمة الافصاح :

يجب التحقق من المستجيب لتحديد نسبة الاشخاص المتأكد من انتماءهم للمجموعة المخفية حيث يتم اضافة سؤال كم عدد الاشخاص الذين ابلغوك او انك قد رأيتهم يعملون شيء يجعلهم ضمن مجموعة مخفية معينة ، فلو رمزنا لهذا التحيز بالرمز τ_K وهو يمثل نسبة الاستجابة عن المجموعة K ، وهنا سنفرض ان قيمة τ_K مساوية الى 1 في المجاميع المعلومة والتي عددها $(K-1)$ وتكون القيمة اقل او مساوية الى 1 في المجموعة الصعبة الوصول (مثال لو ان 50% من المبحوثين قد صرحوا للمستجيبين انهم مدمنين شيئاً ما فعليه تكون $(\tau_K = 0.5)$ ، ان هذه النسبة تضاف كمضروب في الأنموذج وبذلك يصبح أنموذج الأفصاح المتحيز فيه : [VIII]

$$y_{ik} \sim Binom\left(d_i\ , T_K * {N_K}/{N}\right)$$

$$d_i \sim log\ normal\ (\mu, \sigma^2)$$

ويمكن تحديد القيمة الاولية المضافة بانها تتوزع توزيع بيتا

$$T_K \sim Beta\ (\eta_K\ \ , \nu_K)$$

ان القيم الاولية لكل من (μ, σ, N_K) هي نفسها المستعملة في النماذج السابقة .

في هذا الانموذج يكون τ_K وN_K غير واضحان جدا ويميل الى ان تكون مرتبطة بشكل كبير باللاحق ($posteriori$) ، وانها تعكس بعضهما البعض في سلاسل ماركوف مونتي كارلو $(MCMC)$ ، واذا اعدنا كتابة المعلمات في الأنموذج بحيث ان

$$z_K = \frac{N_K}{\tau_K} \qquad\qquad \omega_K = N_K\, \tau_K$$

وباجراء التعويض نحصل على :

$$\tau_K = \sqrt{\frac{\omega_K}{z_K}} \qquad\qquad N_K = \sqrt{\omega_K\ z_K}$$

وعليه يكون معامل التحويل (Jacobian) مساوياً الى $\frac{1}{2\, z_K}$ ، وعليه يكون التوزيع اللاحق المشترك المشروط $(Joint\ Conditional\ posterior\ distribution)$ لهذا الانموذج بعد اجراء التحويل وفق الافتراضات السابقة وبذلك يصبح التوزيع اللاحق :

$$\pi(\mu,\ \sigma^2, d_i, \omega_K, z_K \mid y_{ik}, N_k, N, \eta_K, \nu_K) =$$

$$\left(\prod_{i=1}^{n}\left[\frac{1}{d_i\ \sigma\sqrt{2\pi}}\ e^{-\frac{(\log(d_i)-\mu)^2}{2\sigma^2}}\ \ \prod_{k=1}^{K-1}\left(\binom{d_i}{y_{ik}}\ \left(\frac{N_k}{N}\right)^{y_{ik}}\left(1-\frac{N_k}{N}\right)^{d_i-y_{ik}}\right)\right]\prod_K\left(\binom{d_i}{y_{ik}}\ \left(\frac{\omega_K}{N}\right)^{y_{ik}}\left(\frac{\omega_K}{N}\right)^{d_i-y_{ik}}\ \frac{1}{\sqrt{\omega_K\ z_K}} * \frac{\sqrt{\frac{\omega_K}{z_K}}^{\eta_K\left(\frac{1}{\nu_K}-1\right)-1}\left(1-\sqrt{\frac{z_K}{\omega_K}}\right)^{(1-\eta_K)\left(\frac{1}{\nu_K}-1\right)-1}}{B\left(\eta_K\left(\frac{1}{\nu_K}-1\right),(1-\eta_K)\left(\frac{1}{\nu_K}-1\right)\right)} * \frac{1}{5} * \frac{1}{7/4} * \frac{1}{2\, z_K}\right.$$

.....(13)

وبما ان معادلة تقدير ω_K والى k من المجموعات الجزئية المعروفة العدد كالاتي :

$$\omega_K = N_K\,\tau_K \quad \ldots.(14)$$

وبافتراض ان $\tau_K = 1$ في المجموعات الجزئية المعروفة العدد (حيث انها تخلو من تحيزات الافصاح) وبذلك تصبح المعادلة رقم (14) كالاتي :

$$\omega_k = N_k$$

وهنا نحدث قيمتي ω_K , z_K بدلا من تحديث N_K , τ_K وبذلك يصبح التوزيع اللاحق المشروط لمعادلة ω_k

$$\ell(\omega_K \mid \mu,\ \sigma^2, d_i , z_K, y_{ik} , N_k , N, \eta_K , v_K) \propto \sum_{i=1}^{n} y_{iK} \log\left(\frac{\omega_K}{N-\omega_K}\right) + \sum_{i=1}^{n} d_i \left(1-\frac{\omega_k}{N}\right) +$$

$$\frac{\eta K\left(\frac{1}{v_K}-1\right)-2}{2} \log \omega_K + \left((1-\eta_K)\left(\frac{1}{v_K}-1\right)-1\right) \log\left(1-\sqrt{\frac{\omega_K}{z_K}}\right) - \log 2\, z_K \ \ldots.(15)$$

اما التوزيع اللاحق الى z_K

$$\ell(z_K \mid \mu,\ \sigma^2, d_i , \omega_K,\ y_{ik} , N_k , N, \eta_K , v_K) \propto$$

$$-\frac{\eta K\left(\frac{1}{v_K}-1\right)}{2} \log z_K + (1-\eta_K)\left(\frac{1}{v_K}-1\right)-1)\ \log\left(1-\sqrt{\frac{\omega_K}{z_K}}\right) + \log 2\, z_K \quad \ldots.(16)$$

ان قيمتي ω_K , z_K يجب ان تكون قيم موجبة وقيمة ω_K اصغر من قيمة z_K ، ان المعلمة ω_K لا يمكنها ان تكون اكبر من مجموع السكان في حين z_K ليس لديها حد اعلى واضح *(upper bound)* ، الا ان $N_K = \sqrt{\omega_K z_K}$ يجب ان تكون اقل من مجموع السكان ويتم تضمين جميع الحدود ذات الصلة ورفض القيم المقترحة لـ ω_K أو z_K اذا كانت لا تقع ضمن الحدود، اما بالنسبة للمعلمات الاخرى فلغرض ضبط تلك المعلمة نستخدم نفس الاجراءات في النماذج السابقه وهي استخدم 2.3 من المرات لباقي الخطأ القياسي الذي يحصل عليه الانحدار من سلسلة صغيرة اولية . [VIII]

أ - <u>تحديث ω_K</u> :-

كما سبق ذكره ان $\omega_K = N_K\,\tau_K$ وهذا يعني ان اكبر قيمة يمكن ان تكون عليها ω_K هي ان تكون مساوية لقيمة N_K وذلك لان τ_K تقع بين (1 , 0) .

في بداية التحديث يتم توليد قيمة بالاعتماد على المعادلة الاتية

$$\omega_K(old) = N_K(start) * \tau_K(start)$$

اذ ان $\tau_K(start)$ تم احتسابها من خلال السؤال الذي تم اضافته الى الاستبانه وهو (كم عدد الاشخاص الذين ابلغوك او انك قد رايتهم يعملون شيء يجعلهم ضمن المجموعة المخفية) ثم يؤخذ هذا المجموع ليقسم على العدد الكلي للمجموعة المخفية .

اما $N_K(start)$ فيتم احتسابها بالاعتماد على اسلوب $(Killworth)$.

ثم يتم توليد قيم لاحقة تتبع التوزيع الطبيعي ذو معلمات محددة بوسط حسابي معتمدا على التوليد السابق و بتباين يكون متاثرا بمعلمة الضبط ، ان القيم المتولدة يجب ان تقع ضمن فترة حدود مشروطه تفرض عليها اما اذا فشلت في تحقيق جميع هذه الشروط فاننا سوف نعتمد القيمة المتولدة اولا باسلوب $(Killworth)$ وهذه الشروط هي :

- $\omega_K(new) < \mathrm{N}$
- $\omega_K(new) > 0$

- $\frac{\omega_K(new)}{z_K} < 1$
- $\sqrt{(\ \omega_K(new) * z_K} < \mathrm{N}$

ثم نجري نفس خطوات ميتروبلس $(Metropolis)$ التي اجريت على المعلمات الاخرى في النماذج التي سبق ذكرها .

ب - تحديث z_K :-

يتم التحديث باسلوب مشابه تماما لتحديث ω_K الا انه يختلف عنه في الشروط الموضوعة والتي هي :

- $z_K(new) > 0$
- $\frac{\omega_K}{z_K(new)} < 1$
- $\sqrt{(\ \omega_K(new) * z_K} < \mathrm{N}$

والاختلاف الاخر هو التعويض في المعادلة رقم (16) بدلا من المعادلة رقم (15)

رابعا- الأنموذج المشترك أو المركب $(Combined\ model)$:-

في هذا الانموذج سوف تظهر تأثيرات الانموذجين السابقين ($barrier\ \&\ transmission$) ولغرض توليد تقديرات غير متحيزة يجب معالجة تحيزات المصدرين معا ، ان y_{ik} يتوزع كالاتي:

$$y_{ik} \sim Binom(d_i\ , \tau_K\ * q_{ik}\ )$$
$$q_{ik} \sim Beta\,(\,m_k\ , p_k\ )$$
$$di \sim log\,normal\,(\,\mu,\ \sigma^2\,)$$

ان المعلمات $(m_K\ , \rho_k\ , \rho_{ik}\ , \tau_K)$ جميعها موضوعه تحت قيد وهو ان الفترة تكون ضمن (0 ، 1) .

في هذا الأنموذج يعتمد نفس التحديث السابق بالنسبة للـ $(\ \mu\ , \sigma\ )$، بينما المعلمات الاخرى $(m_K\ , \rho_k\ , d_i)$ يكون التحديث مقارب جدا لتلك الموجودة في النماذج السابقة حيث يكون الفرق هو التعويض بالتوزيع اللاحق لكل معلمة.

ان القيم الاولية المستخدمة هنا هي نفسها المستخدمة في النماذج السابقة اما التوزيع اللاحق لهذا الأنموذج يكون كالاتي :

$$\ell(\mu, \sigma, d_i, m_K, \rho_k, q_{ik}, \tau_k \mid y_{ik}, m_k N, \eta_K, \nu_K) =$$

$$\left(\prod_{i=1}^{n}\left[\frac{1}{d_i\ \sigma\sqrt{2\pi}}\ e^{-\frac{(\log(d_i)-\mu)^2}{2\sigma^2}}\ \prod_{k=1}^{K}\left(\binom{d_i}{y_{ik}}\ (\tau_k\ q_{ik})^{y_{ik}}\ (1-\tau_k\ q_{ik})^{d_i-y_{ik}}\right)\ \frac{q_{ik}^{m_k\left(\frac{1}{\rho_k}-1\right)-1}(1-q_{ik})^{(1-m_k)\left(\frac{1}{\rho_k}-1\right)-1}}{B\left(m_k\left(\frac{1}{\rho_k}-1\right),\ (1-m_k)\left(\frac{1}{\rho_k}-1\right)\right)}\right]\prod_K \frac{\tau_K^{\eta_K\left(\frac{1}{\nu_K}-1\right)-1}(1-\tau_K)^{(1-\eta_K)\left(\frac{1}{\nu_K}-1\right)-1}}{B\left(\eta_K\left(\frac{1}{\nu_K}-1\right),\ (1-\eta_K)\left(\frac{1}{\nu_K}-1\right)\right)} * \frac{1}{m_K} * \frac{1}{5} * \frac{1}{\frac{7}{4}} \quad \dots(17)$$

اذ ان (η_K, ν_K) في توزيع τ_k سيحددان بناءً على معلومات خارجية . وهنا لا يمكن استخدام توزيع بيتا - بانوميل في دمج q_{ik} ويرجع ذلك الى وجود τ_k في الأنموذج ، ومن ثم يجب اخذ عينات من q_{ik} وكذلك زيادة كبيرة في عدد معلمات الأنموذج

أ - <u>تحديث τ_K :-</u>

هذا المؤشر عبارة عن متجه ابعاده (عدد المجموعات المعروفة العدد + المجموعة غير المعلومة ، 1) وبما ان المشكلة تكمن في الافصاح عن الانتماء للمجموعة غير المعروفة العدد فعليه تكون بيانات المتجه بعدد $(k-1)$ مساوية الى 1 بينما القيمة الاخيرة تكون اصغر او مساوية الى 1 والتي تم احتسابها عن طريق مؤشر في الاستبانة وهذه القيمة هي التي سوف تحدث ويكون شكل المتجه كالاتي:

$$tau.k = [\,1\ 1\ \dots\dots\dots\dots\dots 1\ \tau_K]$$

ولتحديث τ_K يكون التوزيع اللاحق كالاتي :

$$\ell\left(\tau_K \,\middle|\, y_{ik}, N, d_i, \mu, \sigma, m_k, m_K, q_{ik}, \rho_k, \eta_K, \nu_K\right) \propto$$

$$\sum_{i=1}^{n}\left(y_{iK}+\eta_K\left(\frac{1}{\nu_K}-1\right)-1\right)\log(\tau_K) + \sum_{i=1}^{n}(d_i-y_{iK})\log(1-\tau_k q_{ik}) + n\left((1-\eta_K)\left(\frac{1}{\nu_K}-1\right)-1\right)\log(1-\tau_K) \quad \dots(18)$$

ومرة اخرى يستخدم انعكاس التماثل الاعتيادي لتلك القيم التي تقع خارج حدود الفترة ، وكذلك يستخدم 2.3 كمعلمة ضبط

ضمن خطوات ميتروبلس $(Metropolis)$ التي اجريت على المعلمات السابقة في النماذج التي سبق ذكرها.

ب- <u>تحديث q :-</u>

q : هو عبارة عن مصفوفة بابعاد $(n * K)$ احد ابعادها يكون مساويا لعدد المستجيبين بينما البعد الاخرى يكون بعدد المجموعات الجزئية المعروفة العدد وغير المعروفة العدد، وتقع جميع القيم ضمن الفترة (0 ، 1).

ولتحديث q_{ik} يكون التوزيع اللاحق الشرطي كالاتي :

$$\ell\left(q_{ik} \,\middle|\, y_{ik}, N, d_i, \mu, \sigma, m_k, m_K, \rho_k, \tau_k, \eta_K, \nu_K\right) \propto \left(y_{ik} + m_k\left(\frac{1}{\rho_k}-1\right)-1\right)\log(q_{ik}) + (d_i - y_{ik})\log(1-\tau_k q_{ik}) + \left((1-m_k)\left(\frac{1}{\rho_k}-1\right)-1\right)\log(1-q_{ik}) \quad \dots\dots(19)$$

وبنفس الافتراضات السابقة على بعض المتغيرات السابقة وهو الارتداد المتماثل $(symmetric\ bounce\ back)$ ، وبه نتمكن من معالجة القيم التي تقع خارج الفترة المحددة . وكذلك يستخدم 2.3 كمعلمة ضبط ضمن خطوات ميتروبولس $(Metropolis)$ التي اجريت على المعلمات السابقة في النماذج التي سبق ذكرها .

التطبيق والنتائج :

أ- جمع البيانات (Data Collection):

بالنظر لافتقارنا لبيانات اي مجموعة من المجموعات المخفية (مثلا متعاطي الحبوب المخدرة) بسبب عدم التوجه بإقامة مسوحات ذات توجهات تعنى بتلك المجاميع مما صعب على الباحثين في كافة الاختصاصات التوجه في هذا المجال مما جعله مهملاً واوصد الابواب لايجاد تقديرات موثوقة .

وكذلك يلعب الجانب الأمني دوراً سلبياً في الحصول على بيانات عن تلك المجاميع مما يجعلها من اهم المعوقات التي تواجه الباحثين في مثل هكذا توجهات. لذلك تم الشروع بتصميم استبانة مخصصة لهذا الغرض بعد ان تم الاطلاع على تجارب الامم المتحدة في اقامة مثل هكذا مسوحات والدراسات المماثلة والتي تجرى في العديد من الدول ، اذ تتخلف بعض التجارب من دولة لاخرى حسب توفر البيانات الموجودة و طبيعة المجتمع المدروس والدراسة المطلوبة الا انها جميعها تتفق في الشكل العام للاستبانة والذي هو عادة ما يتكون من المعلومات العامة للمستجيب كالعمر والتحصيل العلمي وغيرها. اما الجزء الاهم في الاستبانة فيتضمن الاسئلة عن المجتمعات المعروفة وغير المعروفة العدد والتي يطلب من المستجيب تحديد عدد الافراد الذين يعرفهم من تلك المجاميع .

تم اختيار 20 مجموعة جزئية معروفة العدد وكانت نسبة المجموع الكلي لافراد تلك المجاميع تجاوز 20% من حجم السكان بينما كانت هنالك مجموعة جزئية واحدة غير معروفة العدد وهي اعداد متعاطو الحبوب المخدرة .

ولقد اشترط الباحث ان اقصى حد من افراد الشبكة الشخصية للمستجيب هو 30 فردا للمجموعة الواحدة ان السبب وراء وضع قيمة كحد أعلى تحسبا ان يكون المستجيب ضمن هذه المجموعة فعندما يكون المستجيب تدريسياً فمن البديهي ان تكون اجابته عن سؤال كم عدد التدريسيين الذين تعرفهم عالية جدا مقارنة باشخاص يزالون مهنه اخرى وقد تتجاوز 30 شخصا مما سوف يسبب المبالغة بالتقدير $(overestimated)$ في هذه المجموعة، علماً بأن اسلوب القطع لا يؤثر بنسبة كبيرة على البيانات وهذا ما ذكره $(McCormick)$ وآخرون ان هذا القطع سوف يؤثر على البيانات بنسبة ضئيلة تقدر ب 0.25%[X] ، اما عندما يكون القطع عند 50 شخصا فانه لا يؤثر تماماً [XVI] الا انه تم الاشارة في نفس المصدر على ضرورة ان يكون القطع عند 30 شخصاً .

ب- تصميم العينة (Sampling)

بالنظر لعدم وجود بيانات لعدد متعاطي الحبوب المخدرة وكذلك عدم وجود اطر باسماء المتعاطين فان الاسلوب المستعمل في تحديد حجم العينة يختلف من بلد لاخر ومن دراسة لاخرى، ولقد تم حصر العينة على الذكور فمن الصعوبة اجراء مقابلة مع النساء وسؤالهن عن عدد المتعاطين للحبوب .

ان تحديد المجتمع يختلف من دراسة الى اخرى فلو كانت المجموعة المستهدفة هم الاشخاص المصابين بمرض العوز المناعي (AIDS) لكان من الضروري شمول النساء بالعينة وبالتالي تكون النتائج على مستوى الجنسين بينما هنا ستعكس النتائج للذكور فقط ، تم استبعاد الاشخاص دون سن 15 لصعوبة فهمهم لبعض محتويات الاستبانة ، لهذا حدد الباحث الاطار العام للدراسة من عمر 15 سنة ولغاية 60 سنة للذكور من سكنة محافظة بغداد حصراً .

ولغرض تقدير حجم العينة تم استعمال صيغة $(Yamane)$ [VI]

$$n = \frac{N}{1 + N * e^2}$$

اذ n : حجم العينة المطلوبة

N : حجم المجتمع والبالغ (2,331,203) فردا حسب بيانات وزارة التخطيط / الجهاز المركزي للاحصاء لعام 2016

e : مستوى الخطأ والبالغ 0.1 اي ان مستوى الثقة يبلغ 90%

$$n = \frac{2,331,203}{1 + 2,331,203 * 0.1^2} = 99$$

ويتم اضافة ± 5 كعدم استجابة وبذلك تصبح العينة الكلية 104 شخص

ج- العمل الميداني :

تم الشروع بعملية جمع البيانات في مستشفى ابن رشد التعليمي/ بغداد في شهر نيسان لغاية شهر حزيران من سنة 2017 حيث تم مقابلة 104 شخص وجها لوجه من مراجعين او راقدين او احد ذويهم للتخلص من الادمان الدوائي

د- النتائج :

لغرض الحصول على النتائج النهائية تم استعمال حزمة ($NSUM$) الخاصة بتحليل النماذج واستخراج النتائج وهذا الحزمة موجودة ضمن برنامج R النسخة (3.4.1) والمتوفرة على شبكة الانترنيت ، فيتم تقدير المجموعات المعروفة العدد والذي يسمى اسلوب التقدير العكسي ($Recursive\ Back$) ، ثم ايجاد نسبة القيمة المقدرة الى القيمة الحقيقية اذ يتم قبول المجموعة التي تكون نسبتها ضمن المدة بين النصف والضعف (0.5 – 2) بينما ترفض المجموعة التي تكون خارج هذا المدى حيث يكون التقدير اما مبالغا فيه $(overestimated)$ اذ يتجاوز التقدير ضعف القيمة الحقيقية ، او ان يكون التقدير اقل من القيمة الحقيقية $(underestimated)$ فكلما كان افراد المجموعة كبيرة جداً كلما انخفض مقدار التوقع وذلك لصعوبة تذكر كل الافراد الذي تعرفهم ضمن هذه المجاميع ، ان الغاية الرئيسية من تشذيب المجموعات هو لتحسين قيمة تقدير حجم الشبكة الشخصية ، وبعد اجراء عملية التقدير العكسي تبين ان هنالك 12 مجموعة معروفة العدد من اصل 20 مجموعة مختارة يمكن الاعتماد عليها في التقديرات اللاحقة .

بعد ذلك يتم تقدير المجموعة المخفية فتبين ان هنالك 33,183 شخص يتناولون الحبوب المخدرة من سكان بغداد من الذكور والذين يتراوح اعمارهم (15-60) سنة حسب انموذج $(Killworth)$(والذي يعد الانموذج القياسي والمعتمد في اغلب دراسات الامم المتحدة في تقديرات الامم المتحدة [XIII] ، الا ان هذا الانموذج يعاني من بعض التحيزات لذا تم اللجوء الى طرائق جديدة لمعالجتها.

اما التقدير بالاعتماد على النماذج البيزية فاسلوب عملها يعتمد على توليد مجموعة من العينات بحجم 10,000 عينة على ان يتم حرق اول 100 مجموعة منها وبتكرار مقداره 10,000 ، ولغرض المقارنة بين هذه النماذج يتم حساب كفاءة تقدير كل انموذج من نماذج بيز نسبة الى تقدير $(Killworth)$، ويحسب بموجب الصيغة الاتية :

$$efficiency\ of\ estimate = \frac{var\ in\ bayesian\ model}{var\ in\ Killworth\ model}$$

فاذا كانت نتيجة المقارنة مساوية الى الواحد الصحيح دل ذلك على تساوي الطريقتين في التقدير ، اما اذا كانت النتيجة أقل من واحد صحيح يعني ان طريقة بيز اكثر كفاءة من تقدير $(Killworth)$ ، اما للمقارنة بين طرائق بيز تكون افضلها هي اقلها التي تتمتع بأقل كفاءة نسبية .

والجدول رقم (1) يوضح التقديرات الناتجة لاعداد مدمني الحبوب المخدرة وفق الطرائق المستعملة و نلاحظ ان انموذج الدرجة العشوائية $(\mathbf{degree})$ هو اقلهم من ناحية الانحراف المعياري و الكفاءة النسبية بين النماذج المستعملة في التقدير ، وهذا يعني انه افضل النماذج البيزية المستخدمة لهذه البيانات .

ونلاحظ ايضا ان متوسطات النماذج جميعها تقع ضمن فترة القيم الاولية (3 , 8) بينما الانحراف المعياري يقع ضمن الفترة الاولية (0.25 , 2) .

الجدول رقم (1)
التقدير والمتوسط والانحراف المعياري والكفاءة النسبية لتقدير اعداد مدمني الحبوب المخدرة حسب طرائق التقدير

indicate	Killworth	degree	barrier	transmission	combined
NK	33,183	32,862	4,267	57,388	1,570,921
mean	3.8481	3.8822	5.9949	3.8823	5.7914
sigma	0.4102	0.3427	0.3532	0.3434	0.5412
efficiency		0.6980	0.7414	0.7008	1.7407

المحاكاة (Simulation) :

ان السبب الرئيسي من اجراء المحاكاة هو للمقارنة بين النماذج ، ولغرض اجراء المحاكاة نقوم بتوليد اربعة مستويات من البيانات وكل مستوى يحتوي على مختلف من التأثيرات والتحيزات ، حيث يكون اول مستوى يخلو من التأثيرات والتحيزات اما المستوى الثاني يحتوي على تأثيرات الميول $(Barrier\ effects)$ بينما المستوى الثالث يكون متأثر بالافصاح المتحيز $(Transmission\ bias)$ ، اما المستوى الاخير فيضم التأثيران معا .

في المستوى الاولى تكون البيانات تتبع افتراض أنموذج الدرجة العشوائي وهو ان دقة المستجيبين تتبع توزيع $(log\ normal)$ بينما اعداد الاشخاص المعروفين في كل مجموعة يتبع توزيع $(binomial)$ استنادا الى دقة المستجيبين ونسبة مجموع الافراد في كل مجموعه الى السكان الكلي ، اما عند توليد البيانات التي تحتوي على تأثيرات الميول فاننا نضيف تأثير بيتا العشوائي $(beta\ random)$ الى نسبة $(binomial)$ ، اما عند توليد بيانات تعاني من تحيز في الافصاح فاننا هنا نضيف مضاعف τ_K الى نسبة $(binomial)$. وفي الحالة الاخيرة وهي توليد بيانات متأثرة بكلا التحيزين فاننا نضيف تأثير $(beta\ random)$ و مضاعف τ_K الى نسبة (binomial) .

في جميع الحالات التي ذكرت فاننا سوف نستخدم حجم عينة بمقدار (10,000) مجموعة مع حرق اول 100 مجموعة منها وسوف نحاكي بتكرار مقداره (10,000)، في كل مستوى من البيانات المتولدة من المحاكاة سوف نطبق انموذج (Killworth) لغرض اعتماده في ايجاد الكفاءة النسبية لكل طريقة من طرائق بيز المستعملة .

وسيتم اعتماد تقدير حجم اعداد المتعاطين والبالغ (33,183) شخصا والذي تم الحصول عليه سابقا بطريقة $(Killworth)$ كنقطة انطلاق .

نتائج المحاكاة :-

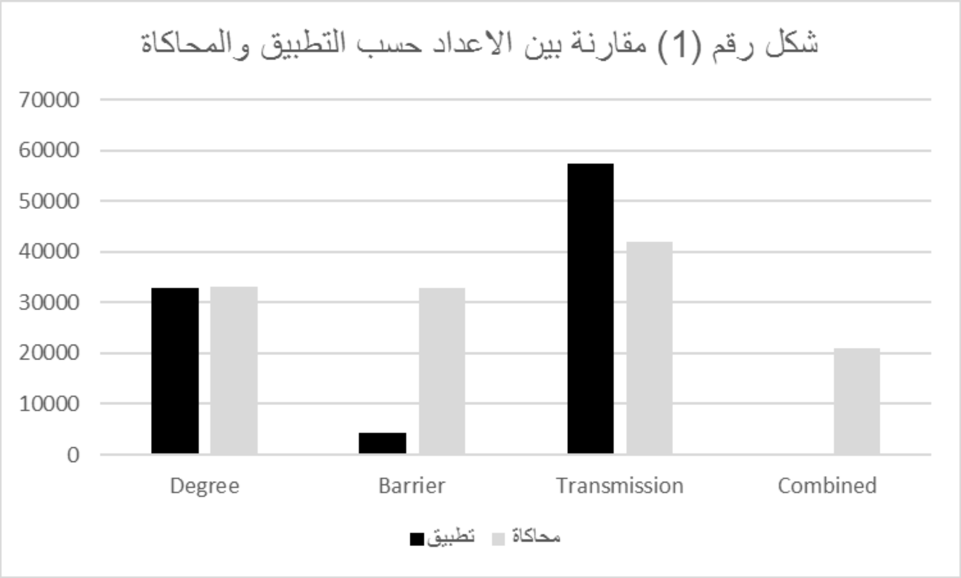
تشير نتائج المحاكاة الى افضلية انموذج الدرجة العشوائي $(degree)$ على نماذج بيز الاخرى فنلاحظ انه يتمتع بأقل انحراف معياري ، لاحظ الجدول (2)

الجدول رقم (2)

التقدير والانحراف المعياري والكفاءة النسبية بالاعتماد على نتائج المحاكاة لتقدير اعداد مدمني الحبوب المخدرة حسب طرائق التقدير

sim \ Model		degree	barrier	Transmission	combined
estimate	N	33,175	32,762	42,031	21,025
	sigma	1.0166	1.0173	1.0259	1.1997
Killworth	N	33,708	36,951	20,304	20,147
	sigma	1.0256	1.0137	1.0070	1.1469
efficiency		0.9825	1.0071	1.0379	1.0942

الشكل رقم (1) يوضح نتائج التقدير حسب نتائج البيانات والمحاكاة ، وفيه تم حذف نتيجة التقدير للانموذج المشترك بسبب صعوبة الرسم في ظل تواجدها بسبب تقديرها المبالغ فيه ، وهنا نلاحظ ان انموذج الدرجة العشوائي ان نتائجه متقاربة في حالة التطبيق والمحاكاة بينما هنالك تفاوت في النماذج الاخرى .



3 : الاستنتاجات

1- ان افضل الطرائق البيزية هو أنموذج الدرجة العشوائي فقد لوحظ انه يتمتع باقل كفاءة نسبية، وان نتائج المحاكاة كانت داعمة لهذا الاستنتاج ، وهذا لا يعني ان هذا الانموذج هو الافضل على الاطلاق .

2- ان الاقرار بأن أنموذج الدرجة العشوائية هو الافضل بين النماذج البيزية يدل على ان البيانات لم تكن تعاني من اي تحيزات، ومن ثم نستنتج ان اسلوب جمع البيانات يكون افضل اذا اجريت المقابلات مع افراد ينتمون للمجموعة المستهدفة.

3- بلغ اعداد المتعاطين للحبوب المخدرة (32,862) شخص متعاطي من الذكور في بغداد للفئة (15-60) وفق انموذج الدرجة العشوائي وهو مقارب بالنتيجة لاسلوب ($Killworth$) القياسي المعتمد والذي بلغ (33,183) شخصاً .

التوصيات

1- بناء أنموذج بيزي يحاول معالجة مشكلة تحيز التذكر لدى المستجيبين ($recall\ bias$) .

2- استخدام توزيعات اخرى في تقدير حجم الشبكة الشخصية ابرزها توزيع بواسون ، اما لاثبات دقة اختيار التوزيع فيتم تقدير المجتمعات المعروفة العدد فاذا كانت تقع ضمن المدى (0 .5 - 2) فيمكن الاستمرار بهذا النهج.

3- يجب ان لا تقل اعداد المجاميع الجزئية المعروفة العدد عن 20 مجموعة جزئية.

Reference

1- Bernard، H،Russel & et al (1989) “Estimating the Size of an Average Personal Network and of an Event Subpopulation”، in The small World، ed. Kochen، M.، New Jersey: Ablex Press ، pp (159-175).

2- Casella، George. George، Edward I. (Aug ،1992) “Explaining the Gibbs Sampler “، The American Statistician ، Volume 46، Issue 3 ، pp(167-174).

3- Gelman، Andrew & Rubin Donald B. (1992) “Inference from Iterative Simulation Using Multiple Sequences”، Statistical Science Vol.7 No. 4 pp(457-511).

4- Habecker، Patrick. Dombrowski، Kirk & Khan، Bilal (2015) “Improving the Network Scale-Up Estimator: Incorporating Mean of Sums، Recursive Back Estimation، and Sampling Weights، PLOS ONE/ journal. Pone.014306، pp(1-16).

5- Jackson،Daniel & et al (2005)، “Social Network Analysis and Estimating the Size of Hard-to-Count Subpopulation”، INSNA، pp (49-60). 4.

6- Jan Mora, Robert & Kloet, Bas (2010) “Digital forensic sampling”, Company: Hoffmann Investigations, Almere The Netherlands , pp (1-9) http://en.hoffmannbv.nl.

7- Kiev International Institute Of Sociology (2009) “Estimating the Size of Populations with High Risk for HIV Using the Network Scale-Up Method” Ukraine.

8- Maltiel، Rachael. & et al (2015) “Estimating Population Size Using The Network Scale-Up Method”، The annals of Applied Statistics، pp (1247-1277).

9- McCarty،Christopher. Killworth،Peter D. Bernard،Russel H. (2001)، “Comparing Two Method for Estimating Network Size” ،Human Organization ،pp(28-39).

10- McCormick، Tyler H. Salganik، Matthew J. Zheng، Tian (2010) "How Many People Do You Know?: Efficiently Estimating Personal Network Size" ، Journal of the American Statistical Association، pp(59-70).

11- Raftery ، Adrian E. Lewis، Steven M. (1996)، "Implementing MCMC"، in Markov Chain Monte Carlo in practice، eds. W.R. Gilks، D. S. and Richardson، S. London: Chapman and Hall، pp(115-130).

12- Syversveen، Anne، R. "Non informative Bayesian Priors. Interpretation and Problems with Construction and Application، pp(1-11)
https://www.ime.unicamp.br/~veronica/MI402/Randi21998.pdf

13- UNAIDS & the US office the Global AIDS Coordinator (2012) "Consultation on Network scale-up and other size estimation methods from general population surveys" ، Kelsey Case، London.

14- Wang، J. et al، (2015) ، "Application of Network Scale Up Method in the Estimation of Population Size for Men Who Have Sex with Men in Shanghai، China " ، PLOS ONE/DOI:10.1371/journal.pone.0143118، pp(1-12).

15- Warwick،J. (2004)، "A data-base method for selecting tuning parameters in minimum distance estimators"، ELSEVIER، COMPUTATIONAL STATISTICS& DATA ANALYSIS، pp(571-584).

16- -Zheng، Tian. Salganik، Mathew J. Gelman، Andrew. (2006)، "How Many People Do You Know in Prison"، Journal of the American Statistical Association، pp(409-423)

Bayesian methods to estimate sub - population

Abstract

The aim of the research is to estimate the hidden population. Here، the number of drug users in Baghdad was calculated for the male age group (15-60) years old ، based on the Bayesian models. These models are used to treat some of the bias in the Killworth method Accredited in many countries of the world.
Four models were used: random degree، Barrier effects، Transmission bias، the first model being random، an extension of the Killworth model، adding random effects such as variance and uncertainty Through the size of the personal network، and when expanded by adding the fact that the respondents have different tendencies، the mixture of non-random variables with random to produce the model of the effects of tendencies، the other extension of the model of the degree is to add the lack of awareness of respondents to produce a model of biased Transmission bias، and for the purpose of improving it by adding other information، The composite or common model that combines the second and third methods .

The use of R software version 3.4.1 which is available on the Internet، was used. The random degree is the best model according to the data collected according to statistical questionnaire prepared for this purpose. And the results of the simulation were supportive of this conclusion، while the number of abusers according to this method 32،862 people abusers.

Keywords/ NSUM, Hidden population, (Scale-up , Killworth, barrier & transmission's methods)