مجله تخصصی اپیدمیولوژی ایران؛ 1393، دوره 10، شماره 2: صفحات 70-62.

مقاله مروری

مقدمه‌ای بر مطالعه‌های صید - بازصید

جعفر حسن‌زاده[1]، مهشید ناصحی[2]، عبدالرضا رجایی‌فرد[1]، دائم روشنی[3]، ابراهیم قادری[3]

[1] گروه اپیدمیولوژی، دانشکده بهداشت و تغذیه، دانشگاه علوم پزشکی شیراز، شیراز، ایران

[2] گروه اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی تهران، تهران، ایران

[3] گروه اپیدمیولوژی و آمار، دانشکده پزشکی، دانشگاه علوم پزشکی کردستان، سنندج، ایران

نویسنده رابط: دائم روشنی، نشانی: سنندج، خیابان پاسداران، دانشگاه علوم پزشکی کردستان، دانشکده پزشکی، تلفن: 087-31827360

پست الکترونیکی: d.roshani@muk.ac.ir

تاریخ دریافت: 92/4/28؛ پذیرش: 92/6/2

مقدمه و اهداف: به تازگی مطالعه‌های صید- بازصید در پژوهش‌های بهداشتی به وفور مورد استفاده قرار می‌گیرد و پژوهشگران به استفاده از این‌گونه مطالعه‌ها در مسایل بهداشتی تمایل زیادی پیدا کرده‌اند. در این مقاله، اصول اولیه و نکات مهمی که در طراحی این مطالعه‌ها باید مدنظر قرار بگیرد، بحث شده است. در ابتدا توضیح مختصری در مورد مطالعه‌های صید- بازصید ارایه می‌شود، سپس پیش‌فرض‌های این نوع مطالعه‌ها مورد بحث قرار می‌گیرد. نکاتی که در این مقاله ارایه می‌گردد، بر اساس پیش‌فرض بسته بودن جامعه مورد مطالعه است و بر همین اساس، چگونگی محاسبه‌ها ارایه شده است. فرمول‌های محاسباتی بر اساس دو صید ارایه شده است و تأثیر وابستگی بین صیدها مورد بحث قرار گرفت. سپس انجام مطالعه‌های صید- بازصید با استفاده از چندین فهرست بحث شده است.

واژگان کلیدی: اپیدمیولوژی، صید- بازصید، نمونه‌گیری

مقدمه

در حالتی که نیاز به برآورد تعداد واحدهای یک جامعه باشد، اما امکان سرشماری در آن جامعه موجود نباشد، می‌توان از روش‌های صید- بازصید استفاده کرد. استفاده از این روش به سال 1896 میلادی باز می‌گردد که پترسون برای برآورد جمعیت ماهی‌های یک منطقه استفاده نمود. در سال 1930 میلادی نیز لینکولن کار مشابهی انجام داد. بعدها در سال 1949 میلادی در هند نیز این روش به کار گرفته شد تا برآوردی از مرگ و تولد در یک منطقه به دست آید (1). این مطالعه در صورت استفاده از دو فهرست dual systems estimation و در صورت استفاده از بیش‌تر از دو فهرست multiple systems estimation نیز گفته می‌شود. در حال حاضر این‌گونه مطالعه‌ها، توسعه‌ی زیادی یافته‌اند (3-2). در حال حاضر مطالعه‌های متعددی با این روش در ایران نیز انجام شده است (7-4).

ایده اولیه این مطالعه‌ها بسیار ساده است. ابتدا یک نمونه‌گیری تصادفی در یک جامعه انجام می‌شود و نمونه‌ها علامت‌دار و به جامعه برگردانده می‌شوند. فرض بر این است که این نمونه‌های علامت‌گذاری شده به صورت آزاد بین سایر واحدهای نمونه‌گیری پخش می‌شوند. سپس با انجام نمونه‌گیری تصادفی دیگر و از روی نسبت افراد دارای علامت در نمونه دوم، تعداد کل جامعه برآورد می‌شود (آزمایش پترسن[1])

مطالعه‌های صید- بازصید را می‌توان با روش مستقیم و غیر مستقیم انجام داد، به این ترتیب که در روش مستقیم یک نمونه را انتخاب کرده و علامت‌گذاری کرده و پس از آن بازصید انجام می‌شود. در روش غیر مستقیم از داده‌های موجود مانند فهرست‌های ثبت اطلاعات یا بررسی‌ها استفاده می‌شود. مشاهده یک فرد به عنوان صید در نظر گرفته می‌شود (3)، سپس با استفاده از برآوردگرهایی، برآورد کل جامعه انجام می‌شود.

کاربرد روش‌های صید- بازصید در علم اپیدمیولوژی را می‌توان به سه بخش دسته‌بندی کرد: 1- برآورد تعداد جامعه؛ 2- برآورد میزان شیوع یک حالت یا بیماری بر اساس بررسی‌های انجام شده؛ و 3- بررسی کامل بودن سامانه‌های ثبت اطلاعات (8).

این مطالعه‌ها را با دو دیدگاه جامعه بسته و جامعه باز می‌توان انجام داد: در جامعه بسته فرض بر این است که مرگ‌ومیر، زاد و ولد و مهاجرت وجود ندارد، اما در جامعه‌ی باز، این رخدادها وجود دارد.

[1] Peterson Trial

پیش‌فرض‌ها و نکات مهم در انجام مطالعه‌های صید- بازصید

در مطالعه‌ی صید- بازصید باید به چند نکته توجه کرد تا مطمئن شد که آیا این روش می‌تواند برآورد دقیقی از جمعیت را ارایه دهد:

1- هدف از مطالعه[1]: هدف از مطالعه و چگونگی استفاده از تحلیل آن باید مشخص گردد. لازم است مشخص گردد که آیا کم‌برآورد و بیش‌برآورد می‌تواند اهداف مطالعه را مخدوش کند یا خیر. برای رسیدن به برخی اهداف لازم است برآوردی دقیق انجام شود، اما در برخی موارد کم‌برآوردی یا بیش‌برآوردی هم می‌تواند رضایت‌بخش است.

2- منابع نمونه‌گیری و تعداد آن‌ها[2]: لازم است منابع نمونه‌گیری مشخص شوند و از منابع قابل اعتماد استفاده گردد. همه موارد باید شانس برابری برای صید شدن[3] در هر فهرست را داشته باشند. همه منابع باید طوری انتخاب شوند که هویت موارد مشخص باشد یا قابلیت شناسایی آن‌ها در هر فهرست کاملاً مشخص باشد. از نظر تعداد منابع، بهتر است تعداد منابع نمونه‌گیری حداقل 3 منبع انتخاب گردد، زیرا انتخاب 2 منبع می‌تواند باعث برآوردی اشتباه- کم یا زیاد- شود. استفاده از حداقل 3 منبع اجازه می‌دهد که از تحلیل لگ خطی استفاده شود. می‌توان 5-3 منبع را انتخاب کرد، اما معمولاً انتخاب بیش‌تر از3 منبع تأثیر خاصی روی برآوردها و حدود اطمینان آن‌ها نمی‌گذارد.

3- ارتباط بین منابع نمونه‌گیری یا فهرست‌ها[4]: ارتباط و وابسته بودن فهرست‌ها به هم باید مورد بررسی قرار گیرد، زیرا یکی از پیش‌فرض‌های صید- بازصید عدم وابستگی[5] فهرست‌ها به هم است. وابستگی مثبت بین فهرست‌ها می‌تواند باعث کم‌برآوردی و وابستگی منفی می‌تواند باعث بیش‌برآوردی شود. در حالتی که فهرست‌ها به هم وابسته باشند، می‌توان از ادغام فهرست‌ها یا حذف آن‌ها یا از تحلیل لگ خطی استفاده کرد، تا اثر اریبی[6] آن‌ها خنثی گردد. البته در موارد انسانی این پیش‌فرض همیشه در خطر مخدوش‌شدگی است. مثلاً وقتی در بیمارستان‌ها گزارش‌دهی برخی بیماری‌ها وجود دارد، ارتباط شدیدی بین فهرست بیمارستان‌ها و فهرست موجود در نظام مراقبت بیماری مورد نظر مشاهده می‌شود (3).

4- تعریف مورد[7]: تعریف مورد باید به گونه‌ای باشد که در تمام فهرست‌ها یکسان و دقیق بوده و در طی دوره مطالعه تغییر نکند. تشخیص این موردها باید به یک صورت واحد- مانند بالینی یا آزمایشگاهی- انجام شده باشد. هم‌چنین محدوده جغرافیایی این فهرست‌ها باید مشخص و در یک محدوده‌ی جغرافیایی باشند.

5- صحت تشخیص موارد و طبقه‌بندی بیماران[8]: در حالت ایده‌آل باید ارزش اخباری مثبت هر یک از فهرست‌ها 100 درصد باشد، اما در مطالعه‌های اپیدمیولوژیک نیازی به این شرط نیست. اگر موارد مثبت کاذب در یک لیست زیاد باشد، باعث بیش‌برآوردی روش صید- بازصید خواهد شد. برای کاهش این خطای طبقه‌بندی، با ادغام برخی فهرست‌ها که خطای طبقه‌بندی در آن‌ها زیاد است- مانند ادغام موارد دارای سرطان رکتوم و کولون- می‌توان این خطا را کاهش داد. روش دیگر کاهش این خطا این است که موارد مثبت کاذب را از طریق ارتباط بین داده‌های ثبت شده[9] () بین فهرست‌های دیگر پیدا کرده و حذف کرد. روش آخر این است که با استفاده از ارزش اخباری مثبت در مطالعه‌های دیگر، این خطا را اصلاح کرد.

ارتباط بین داده‌های ثبت شده: این ارتباط یک اصل مهم در مطالعه‌های صید- بازصید است تا از این روش بتوان موارد هم‌پوشانی را مشخص کرد. اگر ارتباط نادرستی بین داده‌های افراد مختلف وجود داشته باشد[10] باعث کم‌برآوردی خواهد شد. اگر موارد تکراری را به خوبی نتوان تشخیص داد و ارتباط موارد یکسان در فهرست‌های مختلف را به درستی نتوان مشخص کرد[11] باعث بیش‌برآوردی می‌گردد. اگر هر دو نوع اریبی هم‌زمان وجود داشته باشد، ممکن است تا حدودی اثر هم را خنثی کنند. شناسایی افراد در فهرست‌های مختلف می‌تواند از راه مشخصات واحدی که برای افراد وجود دارد- مانند شماره ملی و شماره

[1] Purpose and required accuracy of the study
[2] Source-selection and number of sources
[3] Equalcatchability
[4] Relationships between the selected sources
[5] Independence
[6] Bias
[7] Case-definition
[8] Accuracy of diagnosis and disease classification
[9] Record linkage
[10] false-positive links or homonym errors
[11] false-negative links or synonym errors

بیمه- صورت گیرد و در موارد دیگر می‌توان از نام و نام‌خانوادگی، کدپستی، جنسیت، و ... یا ترکیبی از اسم و فامیل و تاریخ تولد استفاده کرد.

در این مطالعه‌ها باید تعداد افرادی که دست‌کم در یک فهرست وجود دارند باید مشخص گردند[1]. لازم است که تعداد موارد تکراری در هر فهرست مشخص شوند. در مواردی که 3 فهرست وجود دارد، با توجه به وابستگی برخی فهرست‌ها به هم، می‌توان همه مدل‌های لگ خطی که در صید- بازصید کاربرد دارند، به همراه درجه آزادی، برازش، حدود اطمینان‌ها و معیار اطلاع[2] تشکیل داد و بهترین مدل با ارایه تفسیر و توجیه را را ارایه داد. همچنین سازگاری درونی[3] مدل لگ خطی باید از طریق مقایسه دو به دو فهرست‌ها و مقایسه یک فهرست با مجموع دو فهرست دیگر صورت گیرد. محدودیت‌هایی در تحلیل صید- بازصید[4] وجود دارد. اگر جمعیت بسته نباشد، مواردی که در یک فهرست یافت می‌شوند را با احتمال بسیار کمی می‌توان در فهرست بعدی مشاهده کرد. این مسأله باعث کاهش احتمال صید واحد مورد مطالعه در فهرست دوم می‌گردد، بنابراین بیش‌برآوردی در محاسبه جمعیت کل رخ می‌دهد و حساسیت نیز کم‌تر از حد معمول محاسبه می‌گردد.

در مدل نمونه‌گیری صید- بازصید داده‌ها برای تحلیل شامل فراوانی مشاهده‌ها در t صید است که برداری از صفر و یک‌هاست، به نحوی که 1 رؤیت و صفر عدم رؤیت هر واحد نمونه‌گیری می‌باشد. پارامترهای مورد نظر وابسته به فرضیه‌های جامعه مورد بررسی از نظر باز و بسته بودن هستند. بر خلاف جوامع بسته، در جوامع باز تولد، مرگ و مهاجرت به داخل و خارج رخ می دهند، بنابراین در جوامع باز، احتمال بقا برابر 1 در نظر گرفته نمی‌شود و چگونگی محاسبات با پیش‌فرض جامعه باز متفاوت می‌باشد.

جوامع بسته:

در جوامع بسته، حجم جامعه در طول آزمایش تغییر نمی‌کند. در جوامع بسته اگر تعداد دفعات نمونه‌گیری (t) برابر 2 باشد و به این گونه فرض شود که N حجم جامعه مورد نظر است؛ X_{11} نمایان‌گر تعداد واحدهای نمونه گیری مشاهده شده در هر دو نمونه؛ X_{12} واحدهای نمونه گیری مشاهده شده در نمونه اول و مشاهده نشده در نمونه دوم؛ X_{21} واحدهای نمونه‌گیری مشاهده شده در نمونه دوم و مشاهده نشده در نمونه اول؛ و X_{1+} و X_{+1} به ترتیب تعداد واحدهای نمونه گیری در نمونه اول و دوم باشند؛ آن‌گاه جدول شماره 1 را می‌توان به صورت زیر رسم کرد (9،3). در این جدول داده‌های فرضی نیز درج شده است تا محاسبه‌های برآوردگرهای مختلف بر اساس آن انجام شود:

جدول شماره1- طرز قرار گیری واحدها در دو نمونه

		نمونه دوم		
		حاضر	غایب	
نمونه اول	حاضر	X_{11} (20)	X_{12} (40)	X_{1+}
	غایب	X_{21} (15)	X_{22} (؟)	
		X_{+1}		N

در این‌جا هدف برآورد مقدار مشاهده نشده X_{22} است که به سادگی منجر به برآورد حجم جامعه یا همان N می‌شود. پس

$$\hat{N} = (X_{11} + X_{12} + X_{21} + X_{22}) = (X_{11} + X_{12} + X_{21} + X_{22}) + \frac{(X_{12} + X_{21})}{(X_{11})}$$

لینکلن و پترسون نشان دادند که برآورد کل جامعه را می‌توان به صورت زیر نشان داد:

$$\hat{N}_{LP} = \frac{X_{1+}.X_{+1}}{X_{11}}$$

$\hat{N}_{LP}$ را برآوردگر لینکلن- پترسون می‌نامند (10). این برآوردگر مقداری اریبی دارد و بیش‌برآورد ایجاد می‌کند، به ویژه در حجم نمونه کم، اریبی آن بیش‌تر خواهد بود. در مواردی که مخرج کسر این برآوردگر برابر صفر باشد از رابطه‌ی زیر که «برآوردگر چپمن» نامیده می‌شود؛ استفاده می‌گردد که اریبی کم‌تری نسبت به برآوردگر قبلی دارد:

$$\hat{N}_{CH} = \frac{(X_{1+} + 1)(X_{+1} + 1)}{(X_{11} + 1)} - 1$$

محاسبه برای داده‌های فرضی بر اساس برآوردگر لینکلن- پترسون، تعداد کل جامعه را برابر 105 نفر برآورد می‌کند؛ یعنی تعداد افرادی که در دو فهرست وجود ندارند؛ برابر 30 نفر خواهد بود. این محاسبه بر اساس برآوردگر چپمن نیز برابر 105 نفر

[1] Case-ascertainment and capture-recapture analysis
[2] Information criteria
[3] Internal consistency
[4] Limitations of capture-recapture analysis in epidemiology

خواهد بود.

برآوردگر دیگر، Chao است (11) که فرمول آن به صورت زیر می‌باشد:

$$\widehat{N}_C = (X_{12} + X_{21} + X_{11}) + \frac{(X_{21} + X_{12})^2}{4.(X_{11})}$$

محاسبه داده‌های فرضی بر اساس این برآوردگر، تعداد کل جامعه را برابر 113 نفر برآورد می‌کند، که تعداد افراد غایب در دو فهرست برابر 38 نفر محاسبه می‌شود.

در صورتی که نمونه‌گیری دوم با جای‌گذاری باشد؛ می‌توان از 3 برآوردگر زیر استفاده کرد (12،13).

برآوردگر دیگر Bailey است. پیش‌فرض این برآوردگر، انجام نمونه‌گیری دوم با جای‌گذاری است. محاسبه ناشی از این برآوردگر هم تفاوت چندانی با برآوردگرهای قبلی نخواهد داشت:

$$\hat{N}_B = \frac{(X_{1+}+1).X_{+1}}{(X_{11}+1)}$$

یکی دیگر از برآوردگرها، برآوردگر Zelterman است که با استفاده از برآوردگر Horvitz–Thompson محاسبه‌ها را بر اساس تقریب Binomial و پوآسون تخمین می‌زند. فرمول زیر برآوردگر زلترمن بر اساس تقریب Binomial است:

$$\hat{N}_Z = \frac{(X_{11}+X_{12}+X_{21})}{1-\left[\frac{(X_{21}+X_{12})}{(X_{21}+X_{12})+2X_{11}}\right]^2}$$

محاسبه برای داده‌های فرضی با این برآوردگر نیز برآوردی برابر 113 نفر را برای کل جامعه ارایه می‌دهد.

همین برآوردگر با استفاده از توزیع پوآسن را می‌توان به صورت زیر نوشت که اثبات این فرمول‌ها را می‌توان در مقاله Brittain یافت (3):

$$\widehat{N_{Zp}} = \frac{(X_{11}+X_{12}+X_{21})}{1-\exp\left(-2\frac{X_{11}}{X_{21}+X_{12}}\right)}$$

محاسبه برای داده‌های فرضی با این برآوردگر، برآوردی برابر 145 نفر را برای کل جامعه ارایه می‌دهد.

برآوردگرهای دیگر بر اساس حداکثر درست‌نمایی و گشتاورها[1] را می‌توانید در مقاله Brittain مطالعه فرمایید(3).

برای محاسبه حدود اطمینان می‌توان از راهنمای زیر استفاده کرد:

1. اگر درصد افراد علامت‌گذاری شده $\left(\frac{X_{11}}{X_{+1}}\right)$ کم‌تر از 10 درصد باشد و

A: X_{11} کم‌تر از 50 باشد، از توزیع پوآسون برای تعیین حدود اطمینان استفاده شود.

B: X_{11} بیش‌تر و مساوی 50 باشد؛ از توزیع نرمال برای تعیین حدود اطمینان استفاده شود.

اگر درصد افراد علامت‌گذاری شده و بازصید شده $\left(\frac{X_{11}}{X_{+1}}\right)$ بیش‌تر یا مساوی 10 درصد باشد؛ برای محاسبه‌ی حدود اطمینان از توزیع دوجمله‌ای استفاده می‌شود.

واریانس این برآوردگر چاپمن توسط Seber ارایه شده است و برابر است با (16-14):

$$Var\widehat{N}_{CH} = \frac{(X_{1+}+1)(X_{+1}+1)(X_{1+}-X_{11})(X_{+1}-X_{11})}{(X_{11}+1)^2(X_{11}+2)}$$

بنابراین فاصله اطمینان (1-α) درصد آن با فرض نرمال بودن برای برآورد $\widehat{N}_{CH}$ برابر است با:

$$\widehat{N}_{CH} \pm Z_{1-\alpha/2}.\sqrt{Var\widehat{N}_{CH}}$$

واریانس برآوردگر Chao و زلترمن نیز با روش زیر محاسبه می‌شود (17):

$$Var\hat{N} = \frac{(X_{21}+X_{12})^2}{4(X_{11})}\left(\frac{(X_{21}+X_{12})}{2(X_{11})}+1\right)^2$$

بنابراین فاصله اطمینان (1-α) درصد با فرض نرمال بودن برای برآورد $\widehat{N}$ برابر است با:

$$\hat{N} \pm Z_{1-\frac{a}{2}}.\sqrt{Var\hat{N}}$$

در این مقاله، فاصله اطمینان با فرض توزیع نرمال مورد بررسی قرار گرفته است، اما بر اساس توزیع پوآسون و دو جمله‌ای نیز قابل محاسبه می‌باشد.

هر چه اریبی به صفر نزدیک‌تر باشد، برآورد درست‌تری به دست می‌آید. این اریبی را می‌توان برای برآوردگرهای مختلف محاسبه کرد و دقت هر کدام را در برآورد جامعه کل، بررسی نمود. به نظر می‌رسد مقدار اریبی در برآوردگر Chao از سایر برآوردگرها کم‌تر

[1] McKendricks Moment Estimator

باشد. مقدار اریبی هر روش با فرمول زیر محاسبه می‌شود (18):

$$Bias = \frac{X_{22}}{(\frac{X_{21}.X_{12}}{X_{11}+1})} - 1$$

وابستگی بین فهرست‌ها:

بین دو فهرست می‌تواند وابستگی مثبت[1] یا وابستگی منفی[2] وجود داشته باشد. در صورت وجود وابستگی مثبت، موارد ثبت شده مشابه در هر دو فهرست زیاد است، یعنی صید یک فرد در یک فهرست باعث افزایش احتمال صید فرد در فهرست دیگر می‌شود. در صورت وجود وابستگی منفی، درصد بیش‌تر افراد در یکی از فهرست‌ها مشاهده می‌شوند و موارد تکراری در دو فهرست بسیار کم خواهد بود، یعنی با صید یک فرد در یک فهرست، احتمال صید در فهرست دیگر کاهش می‌یابد. وابستگی مثبت دو فهرست به هم باعث کم‌برآوردیو وابستگی منفی بین دو فهرست باعث بیش‌برآوردیموارد می‌شود.

در صورتی که چند فهرست وجود داشته باشد، می‌توان برای بررسی وابستگی بین دو فهرست، نسبت شانس را محاسبه کرد و هم‌چنین از آزمون مربع کای استفاده کرد. در صورت معنی‌دار بودن یعنی فهرست‌ها وابسته هستند و در مدل لگ خطی باید اثر متقابل بین آن‌ها در نظر گرفته شود. اگر مقدار نسبت شانس کم‌تر از یک بود، وابستگی منفی-بیش‌برآوردی- و اگر بیش‌تر از یک بود، وابستگی مثبت-کم‌برآوردی- وجود دارد (3).

روش دیگر برای کاهش اثر همبستگی بین فهرست‌ها، روش Wittes است. در این روش، برآورد جامعه را بر اساس ترکیب دو به دو فهرست‌های مختلف بررسی می‌شود. اگر متفاوت بود، می‌توان چند کار را انجام داد: 1- محاسبه نسبت شانس بین فهرست‌ها، و 2- ترکیب فهرست‌های وابسته به هم و انجام مجدد محاسبه‌ها.

حساسیت فهرست‌ها یا صیدها:

برای محاسبه حساسیت هر صید یا فهرست می‌توان از تقسیم تعداد موارد یافت شده در هر فهرست تقسیم بر کل برآورد جامعه استفاده کرد: X_{1+} / N ضرب در 100 و برای روش دوم هم می‌توان گفت X_{+1} / N ضرب در 100. در مطالعه‌های صید-بازصید باید تا حد امکان حساسیت یافتن مورد ها[3] در هر فهرست برابر 100 درصد باشد. هرچند مقدار بالاتر از 75 درصد نیز قابل قبول است.

مثال برای محاسبه حساسیت یک فهرست:

در یک مطالعه، در صید اول تعداد 30 نفر (M) بیمار و در صید دوم تعداد 43 نفر (C) بیمار صید شد که 22 نفر در هر دو صید (R) مشترک بودند. محاسبه با استفاده از فرمول چاپمن به صورت زیر است:

$$\hat{N}_{CH} = \frac{(30+1)(43+1)}{(22+1)} - 1 = 58$$

و حساسیت صید دوم برابر است با:

$$Sensitivity(\%) = \frac{C}{N}\acute{}\ 100 = \frac{43}{58}\acute{}\ 100 = 74\%$$

در مطالعه‌های صید- بازصید روش خاصی برای محاسبه حجم نمونه وجود ندارد، اما به صورت کلی زمانی که چند شرط برقرار باشد، حجم نمونه کافی به نظر می‌رسد.: 1- تعداد افراد مشاهده شده در فهرست اول و دوم بزرگ‌تر از برآورد کل جامعه باشد، که پس از تحلیل مشخص خواهد شد؛ 2- تعداد افراد مشترک در هر دو فهرست بیش‌تر از 7 فرد باشد. در صورت وجود این شرایط، فرض می‌شود که برآورد نااریب خواهد بود.

اختلال در مطالعه‌های صید-بازصید:

در مطالعه‌های صید- بازصید ممکن است مشکلاتی پیش آید که اثرات آن و روش پیش‌گیری در جدول شماره 2 ذکر شده است(19).

نحوه محاسبه با استفاده از لگ خطی (20-21):

مواردی که تا این قسمت ذکر شد، تحلیل ساده‌ای از مطالعه‌های صید- بازصید با استفاده از دو بار صید یا دو فهرست بود. در صورتی که تعداد فهرست‌ها یا صیدها بیش‌تر از دو فهرست باشد؛ استفاده از این محاسبه‌ها ساده به راحتی امکان‌پذیر نیست، و باید از لگ خطی استفاده شود. در این تحلیل، محاسبه‌ها به گونه‌ای انجام می‌شود که اثرات اصلی و تداخلات مختلف بین فهرست‌ها نیز بررسی شود و برآورد تعداد افراد گم شده با هر مدل محاسبه می‌گردد. مدل لگ خطی وابستگی و هتروژنییتی فهرست‌ها را در محاسبه لحاظ می‌کند و در شرایطی که چند منبع وجود داشته باشد، بسیار قوی عمل می‌کند. تداخل بین فهرست‌ها می‌تواند روی مدل تأثیر بگذارد و این تأثیر توسط AB λ در مدل بیان می‌شود. در صورت داشتن سه فهرست یا منبع،

[1] Positive dependence
[2] Negative dependence
[3] case finding

تعداد 8 ترکیب می‌تواند اتفاق افتد:

مقدار تتا، لگاریتم تعداد افراد مورد انتظار است که انتظار می‌رود در همه فهرست‌ها حضور داشته باشند، سه مقدار دیگر مقادیر اثر هر فهرست هستند، سه مقدار بعدی تداخل بین فهرست‌ها است و پارامتر آخر تداخل هر سه فهرست است که البته از طریق داده‌های مطالعه قابل آزمایش کردن نیست و فرض می‌شود که مقدار آن صفر باشد. آزمودن این پارامترها در شرایطی که حجم نمونه کوچک باشد، دارای توان مناسبی نیست. در این مدل‌ها مقدار μ تعداد مورد افراد در جامعه مورد بررسی است.

انتخاب بهترین مدل (25-22):

مدل آماری
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{iK}^{AC}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{jk}^{BC}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{iK}^{AC}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{iK}^{AC} + \lambda_{jk}^{BC}$
$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{iK}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$

انتخاب مدل یکی از مسایلی است که در مطالعه‌های صید- بازصید باید مدنظر قرار گیرد. برای انتخاب مدل چندین راه‌کار وجود دارد. در انتخاب مدل مناسب که بهترین برآورد را دارد، می‌توان از درجه آزادی، مقدار آکائیکه (Akaike Information Criteria(AIC)) و یا مقدار بیزی آکائیکه (Bayesian Information Criterion(BIC)) استفاده کرد. AIC و BIC توسط likelihood ratio tests آزمون می‌گردد و مقدار کم‌تر نشان‌دهنده بهتر بودن برازش است.

آکائیکه زمانی کاربرد دارد که اثر متقابل در مدل وارد شده و هدف انتخاب بهترین مدل باشد، فرمول آن به این صورت است:

AIC = G2 - 2(df)

BIC خود دو نوع متفاوت دارد که Hook and Regal این دو فرم را SID و DIC نامیده‌اند.

SIC = G2 - (ln یا Information Criterion Schwarz Nobs)(df),

DIC = G2 - (ln یاDraper Information Criterion (Nobs/2p))(df),

با استفاده از AIC، معمولاً برازنده‌ترین مدل، مدل اشباع شده- یعنی مدلی با درجه آزادی صفر و k-1 اثرات متقابل- است، اما برآورد جمعیت در این مدل زیاد بوده و حدود اطمینان بازتری را به دست می‌دهد. در مواردی که حجم نمونه کم باشد، دو شاخص دیگر و به ویژه DIC نتایج بهتری را ارایه می‌دهند.

جدول شماره 2- تأثیر پیش‌فرض‌های مختل شده بر محاسبه برآورد جمعیت موارد و حساسیت فهرست‌ها و راه‌کارهای پیش‌گیری از این اختلال‌ها

پیش‌فرض مختل شده	اثر روی تخمین جمعیت موارد	اثر روی حساسیت	روش پیش‌گیری
جامعه بسته	افزایش	کاهش	- کاهش فاصله زمانی نمونه‌گیری از فهرست‌ها - عدم نمونه‌گیری از روزهای خاص-مثلاً خیلی شلوغ یا روزهایی که افراد خاصی مراجعه می‌کنند) - انتخاب دقیق محل نمونه‌گیری - بررسی وضعیت سکونت افراد - تعریف دقیق حدود مرزی در هر محل مطالعه - تهیه نقشه از محیط مطالعه برای صید اول - اطمینان از درک درست جمعیت مورد هدف مطالعه و جلوگیری از اریبی انتخاب - استفاده از افراد آگاه و مطلع برای خارج کردن افراد نا آگاه
احتمال صید برابر یا Equal Catchability	کاهش	افزایش	- غربالگری خانه به خانه برای یافتن بیماران - انتخاب دقیق محل مطالعه - استفاده از روش‌های حساس به بیماران بدحال

یافتن افراد مشابه در فهرست‌ها	True Matched Miss	افزایش	کاهش	- یافتن داده‌های معتبر و کافی که بتواند افراد را به هم وصل کند.
	False Matched created	کاهش	افزایش	- یافتن داده‌های معتبر و کافی که بتواند افراد را به هم وصل کند.
عدم وابستگی فهرست‌ها	وابستگی مثبت	کاهش	افزایش	- استفاده از گروه‌های متفاوت برای جمع‌آوری داده‌ها از هر فهرست یا در هر صید - استفاده از افراد آگاه متفاوت در صید افراد در هر فهرست
	وابستگی منفی	افزایش	کاهش	- ارجاع افراد به سرویس مربوط پس از صید دوم
	کفایت حجم نمونه	افزایش	کاهش	- در صورت امکان افزایش حجم نمونه

نتیجه‌گیری

به صورت کلی، استفاده از این روش نمونه‌گیری در مطالعه‌های بهداشتی در سال‌های اخیر رو به افزایش است. مطالعه‌های صید- بازصید را می‌توان به منظور بررسی وضعیت کامل بودن گزارش‌دهی در نظام‌های مراقبت نیز مورد استفاده قرار داد (26). با استفاده از این مطالعه تعداد مورد انتظار بیماران را می‌توان محاسبه کرد و وضعیت تشخیص و گزارش‌دهی بیماران را بررسی نمود.

این‌گونه مطالعه‌ها دارای نقاط قوت و ضعف هستند، که باید در مرحله طراحی به آن دقت شود. قبل از اجرا، هدف از انجام این مطالعه‌ها دقیقاً مشخص شود تا چگونگی استفاده از آن و تأثیری که اختلال‌های مختلف روی آن می‌گذارد، برای پژوهشگران مشخص باشد. استفاده از روش صید- بازصید با دو فهرست یا دو بار صید، جزء ساده‌ترین روش این مطالعه‌ها است که در این مقاله اصول کلی آن و نکات مهم در مورد آن ارایه گردید. انجام این‌گونه پژوهش‌ها با تعداد صیدهای بیش‌تر یا فهرست‌های بیش‌تر می‌تواند برآوردهای بهتری را به دست بدهد. البته استفاده از بیش‌تر از 3 فهرست یا 3 صید، ارزش افزوده‌ای در این مطالعه‌ها ندارد، و 3 فهرست انتخاب مناسبی است. در بسیاری از مطالعه‌های بهداشتی وابستگی زیادی ممکن است بین فهرست‌ها مشاهده شود، بنابراین در این مواقع می‌توان دو فهرستی که وابستگی زیادی به هم دارند، را در هم ادغام نمود. به طور کلی روش‌های ذکر شده در این مقاله زمانی کاربرد درستی خواهند داشت که سه اصل زیر برقرار باشد:

1- بسته بودن جامعه نسبت به مرگ، تولد، مهاجرت به داخل و خارج؛
2- تمام واحدهای نمونه‌گیری در هر بار نمونه‌گیری دارای شانس یکسان باشند
3- واحدهای نمونه‌گیری، گم نمی‌شوند و توسط افراد مشاهده‌گر به صورت عمدی جستجو نمی‌شوند.

نسبت تعداد افراد مشاهده شده در یک فهرست-یا صید- به افراد مشاهده شده در هر دو فهرست (افراد مشترک) می‌تواند تأثیر عمده‌ای در محاسبه‌ها داشته باشد. در شرایط مختلف، برآوردگرهای ذکر شده می‌توانند برآوردهای متفاوتی را ارایه دهند. این مسأله در مطالعه VAN HEST و همکاران (27) تشریح شده است. اگر نسبت افراد مشاهده شده در یکی از فهرست‌ها به افراد مشاهده شده در هر دو لیست 0/5-1/5 باشد، مدل لگ خطی و برآوردگرهای دیگر نتایج مشابهی را ارایه می‌دهند. اگر تعداد افراد مشاهده شده در یک فهرست بسیار بیش‌تر از افراد مشاهده شده در هر دو فهرست باشد، برآوردهای لگ خطی دچار بیش‌برآوردی خواهند شد (27).

برای انتخاب مدل مناسب و برآورد مناسب، نباید صرفاً بر اساس نوع برآوردگر و خطای محاسبه شده و یا AIC و BIC مدل را انتخاب کرد (8) و باید حتماً توجیه مناسبی نیز داشت. در کل، در این مطالعه‌ها هیچ‌وقت نمی‌توان متوجه شد کدام مدل درست‌ترین است، اما مدل‌های اشتباه را تا حدودی می‌توان تشخیص و خارج کرد.

منابع

1. Herzog TN. Applications of Capture-Recapture Methods. 1, editor: Actuarial Research Clearing House; 2006.
2. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. Stat Med. 2001 Oct 30; 20: 57-123.
3. Brittain S, Böhning D. Estimators in capture–recapture studies with two sources. AStA Adv Stat Anal. 2009;93:23-47.
4. Zemestani AR, Mahmoudi M, Keshtkaar AA, Majdzadeh SR, Foroozanfar MH, Semnani S. Estimation of Cancer Cases in Golestan Province between 2004-2006 by Using Capture-Recapture Method. Medical Journal of Tabriz University of Medical Sciences and Health Services. 2013; 35: 26-33.
5. Khazaei S, Poorolajal J, Mahjub H, Esmailnasab N, Mirzaei M. Estimation of the Frequency of Intravenous Drug Users in Hamadan City, Iran, Using the Capture-recapture Method. Epidemiology and Health. 2012; 34: e2012006.
6. Motevalian SA, Mahmoodi M, Majdzadeh R, Akbari ME. Estimation of death due to road traffic injuries in Kerman district: application of capture-recapture method. Health College and Health institute Journal. 2007;2:61-72.
7. Zavarehac DK, Mohammadia R, Laflammeb L, Naghavide M, Zareief A, Haglunda JA. Estimating road traffic mortality more accurately: Use of the capture–recapture method in the West Azarbaijan Province of Iran. International Journal of Injury Control and Safety Promotion. 2008; 15: 9-17.
8. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev. 1995; 17: 243-64.
9. Herzog TN. Applications of Capture-Recapture Methods.
10. Buckland S, Goudie J, Borchers D. wildlife population assessment, past development and future Directions. Biometrics. 2000; 56: 1-12.
11. Chao A. Estimating the population size for capture-recapture data with unequal catchability. Biometrics. 1987 Dec; 43: 783-91.
12. Chao A, Yip P.S.F, Lee S.M, Chu W. Population size estimation based on estimating functions for closed capture-recapture models. Journal of Statistical Planning and Inference. 2001;92: 213-232.
13. Pollock KH, Nichols JD, Brownie C, Hines JE. Statistical Inference for Capture-Recapture Experiments, Wildlife Monographs: The Wildlife Society; 1990; 107:1-97.
14. Pollock KH. Capture-recapture models: a review of current methods, assumptions and experimental design. Wildl . Monogr. 1981;6: 426-435.
15. Corrao G, Bagnardi V, Vittadini G, Favilli S. Capture-recapture methods to size alcohol related problems in a population. J Epidemiol Community Health. 2000 Aug; 54: 603-10.
16. Seber GA. The effects of trap response on tag recapture estimates. Biometrics. 1970; 26: 13-22.
17. Böhning D. A simple variance formula for population size by conditioning. Stat Methodol. 2008; 5: 410-23.
18. Brittain S, Böhning D. Estimators in capture–recapture studies with two sources. AStA Adv Stat Anal. 2009; 93: 23–47.
19. Notes on using capture-recapture techniques to assess the sensitivity of rapid case-finding methods: VALID International Ltd; 2006.
20. Overstall A, King R. Capture-Recapture Models for Population Estimates: University of St Andrews.
21. Cormack RM. Log-linear models for capture-recapture Biometrics. 1989; 45: 395-413.
22. Akaike H. A new look at the statistical model identification. IEEE Transactions on automatic control. 1974; 19: 716-23.
23. Draper D. Assessment and propogation of model uncertainty. Journal of the Royal Statistical Society, Series B. 1995; 57: 78-9.
24. Schwarz G. Estimating the dimension of a model. The annals of statistics. 1978; 6: 461-4.
25. Hook EB, Regal RR. Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. Am J Epidemiol. 1997 Jun 15; 145: 1138-44.
26. van Hest NA, Smit F, Baars HW, De Vries G, De Haas PE, Westenend PJ, et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? Epidemiol Infect. 2007 Aug; 135: 1021-9.
27. .van Hest NA, Grant AD, Smit F, Story A, Richardus JH. Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data. Epidemiol Infect. 2008 Jan; 136: 14--22.

Iranian Journal of Epidemiology 2014; 10(2): 62-70.

Review Article

Introduction to Capture-Recapture Studies

Hassan Zadeh J[1], Nasehi M[2], Rajaeifard A[1], Roshani D[3], Ghaderi E[3]

1- Dept of Epidemiology, School of Health and Nutrition, Shiraz University of Medical Sciences, Shiraz, Iran
2- Dept of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
3- Dept of Epidemiology and Biostatistics, School of Medicine, Kurdistan University of Medical Sciences, Sanandaj, Iran
Corresponding author: Roshani D., daemroshani@gmail.com

Recently, capture-recapture studies have been used and researchers tend to use these studies in the health field. Therefore, we discussed the basic concepts of these studies. First, we described capture-recapture studies. Then, the important assumptions and calculations were presented according to the close population assumption. Statistical formulas were presented for two-capture methods and dependency between the two lists was discussed. Then, we addressed more than two capture methods.

Keywords: Epidemiology, Capture-Recapture, Sampling