۱

مفاهیم پایه‌ای مبحث نمونه‌گیری

۱.۰ مقدمه

هدف آماردان به‌طور عمده یکی یا چند مورد از موارد زیر است:

الف) وقتی با جامعه‌ای روبه‌روست، یعنی وقتی با متغیری تصادفی مواجه است مدل توزیع آن جامعه یا متغیر را بیابد تا بتواند احتمال رخداد متغیر را در هر بازه‌ای بداند، و در نتیجه با دانستن حجم جامعه، فراوانی متغیر را در آن بازه تعیین کند.

ب) اگر مدل جامعه را می‌داند ولی مدل شامل پارامترهایی مجهول است، برآوردیِ نقطه‌ای و یا بازه‌ای برای آن پارامترها تهیه کند.

ج) وقتی دربارهٔ پارامتر یا مشخصه‌ای از جامعه به‌دلایلی شهودی فرضی دارد آن فرض را در سطح معنادار بودن خاصّی آزمون کند.

برای دستیابی به این هدفها و انجام استنباطها باید از اطلاعات موجود در مشاهدات استفاده کرد و به این دلیل است که قبل از هر چیز و هر اقدام در زمینهٔ انجام استنباط، به‌نمونه‌ای از جامعه نیاز داریم. در این فصل اصطلاحهای فنی مربوط به مسألهٔ نمونه و نمونه‌گیری را تعریف و برخی مفاهیم آماری را که در فرایند اجرای نمونه‌گیریها و کسب نتایج ضروری‌اند مرور می‌کنیم.

۱.۱ عنصر

عنصر، شیء یا موضوعی است که مشخصه‌ای از آن اندازه‌گیری می‌شود.

مثال ۱.۱ در ناحیه‌ای دربارهٔ انتشار یک نوع قرضهٔ ملّی نظرخواهی می‌شود. هدف بررسی، برآورد نسبت موافقین با انتشار قرضه است. در این مثال هر رأی‌دهنده یک عنصر است و اندازه‌ای که برای هر رأی‌دهنده به‌دست می‌آید موافقت یا عدم موافقت او با انتشار قرضه است. ▲

مثال ۲.۱ در خط تولید یک دارو که به‌صورت قرص است معمولاً مادّه مؤثر دارو در هر قرص را کنترل می‌کنند. در این مثال هر قرص یک عنصر است و وزن مادّهٔ مؤثر هر قرص، اندازه‌ای است که از آن قرص نتیجه می‌شود. ▲

۲.۱ جامعه

در هر بررسی آماری، جامعه، گردایهٔ عناصری است که می‌خواهیم دربارهٔ آن استنباطی به‌عمل آوریم. در مثال ۱.۱، جامعه، گردایهٔ همهٔ رأی‌دهندگان ناحیه است و در مثال ۲.۱، جامعه، گردایهٔ همهٔ قرصهای یک پخت داروست. در مسألهٔ نمونه‌گیری معمولاً جامعه متناهی است و تعداد عنصرهای آن را حجم جامعه می‌نامند. کار مهم بررسی‌کننده این است که با دقت و به‌صورتی کامل جامعه را تعریف کند. تعریف جامعه باید به‌صورت توصیف دقیق عنصرهایی باشد که به جامعه تعلق دارند و آن را می‌سازند. به‌علاوه باید نحوهٔ اندازه‌گیری مشخصهٔ عنصرها نیز مشخص شود.

گاهی اوقات مجموعهٔ اندازه‌هایی را که از اندازه‌گیری مشخصه‌ای از عنصرها به‌دست می‌آید جامعه می‌نامند. مثلاً اگر مدیر اتوبوسرانی یک ناحیه مایل به‌اتخاذ روشی برای صرفه‌جویی میزان مصرف بنزین ماهیانهٔ اتوبوسها باشد، اعدادی که معرّف میزان بنزین مصرفی ماهیانهٔ اتوبوسها هستند، جامعه را تشکیل می‌دهند. در همین مثال می‌توان جامعه را گردایهٔ همهٔ اتوبوسهای ناحیه در نظر گرفت و هر کدام از آنها را عنصر جامعه نامید و برای هر عنصر، یعنی هر اتوبوس، اندازهٔ بنزین مصرفی ماهیانه را ثبت کرد.

۳.۱ نمونه

نمونه، بخشی از جامعهٔ تحت بررسی است که با روشی که از پیش معین شده است انتخاب می‌شود، به‌قسمی که می‌توان از این بخش، با توجه به روش انتخاب، استنباطهایی دربارهٔ کل جامعه انجام داد. همان‌طور که متذکر شدیم، در این کتاب روشهای انتخاب نمونه و برخی استنباطهای حاصل از آن را بررسی می‌کنیم. فرایند انتخاب نمونه و استخراج نتایج و استنباطهای حاصل را بررسی نمونه‌ای می‌گویند.

بررسیهای نمونه‌ای را به دو نوع رده‌بندی می‌کنند ــ توصیفی و تحلیلی. در بررسی توصیفی، هدف صرفاً کسب اطلاعاتی دربارهٔ گروههای بزرگ است ــ مثلاً تعداد زنان و کودکانی که برنامهٔ تلویزیونی خاصّی را می‌بینند. در بررسی تحلیلی، بین زیرگروههای متفاوت جامعه، برای کشف تفاوتهای آنها مقایسه‌هایی صورت می‌گیرد و یا فرضهایی را دربارهٔ تفاوتهای موجود بیان کرده و درستی یا نادرستی آنها را آزمون می‌کنند. البته تمایز این دو نوع بررسی نمونه‌ای همیشه کاملاً روشن نیست. در بسیاری از موارد، داده‌ها برای هر دو نوع بررسی به‌کار می‌روند.

۴.۱ واحد نمونه‌گیری

واحدهای نمونه‌گیری گردایه‌های نامتداخل عنصرهایی از جامعه هستند که اجتماع آنها کل جامعه را تشکیل می‌دهند. مثلاً در نمونه‌گیری از افراد یک شهر، هر فرد شهر را می‌توان یک واحد تلقی کرد. در صورت لزوم هر خانوار را می‌توان یک واحد در نظر گرفت، و بالاخره هر بخش شهر را می‌توان یک واحد گرفت. بنابراین صرفاً واحد نمونه‌گیری به تعریف قبلی آمارگر از واحد بستگی دارد. وقتی تعریف خاصی از واحد نمونه‌گیری ارائه نمی‌شود، هر عنصر جامعه را واحد می‌نامند. در سراسر این کتاب منظور ما از واحد نمونه‌گیری همان عنصر است، مگر آنکه قبلاً تعریف خاصی از واحد ارائه شود.

۵.۱ چارچوب نمونه‌گیری

فهرست کامل واحدهای نمونه‌گیری را چارچوب می‌نامند.

اگر در مثال ۱.۱، هر رأی‌دهنده را واحد نمونه‌گیری بنامیم، فهرست نام تمام رأی‌دهندگانی را که به‌عنوان رأی‌دهنده ثبت شده‌اند می‌توان برای چارچوب رأی‌دهندگان به‌کار برد. توجه کنید که این چارچوب شامل همهٔ عناصر جامعه نیست. اگر واحد نمونه‌گیری را خانوار بگیریم آن‌گاه فهرست راهنمای تلفن، یا فهرست همهٔ سرپرستان خانوارها را که از یک سرشماری به‌دست آمده است می‌توان به‌عنوان چارچوب به‌کار برد. تمام این چارچوبها معایبی دارند. فهرستها نمی‌توانند روزانه به‌هنگام شوند. ممکن است نام بسیاری از سرپرستهای خانوارها یادداشت نشده باشد. لذا همیشه تفاوتی بین کل جامعه و چارچوبی که در دست است وجود دارد. باید تا آنجا که ممکن است تلاش کرد تفاوت بین چارچوب و جامعهٔ اصلی، آن قدر کوچک باشد که وقتی نمونه از چارچوب انتخاب می‌شود بر استنباطهای حاصل برای جامعه اثری نگذارد. برای هر مسألهٔ تحت بررسی می‌توان برحسب تعریفی که از واحد نمونه‌گیری می‌شود چارچوبهایی مختلف ساخت. مثلاً در بررسی میزان محصول غلهٔ مزارع یک ناحیه می‌توان واحد نمونه‌گیری را مالک هر ناحیه در نظر گرفت که در این صورت چارچوب نمونه‌گیری فهرست اسامی مالکین مزارع است، که با پرسش از آنها میزان محصول هر مزرعه مشخص می‌شود، و یا می‌توان کل مساحت مزارع را به‌قطعات مثلاً ده هکتاری تقسیم و شماره‌گذاری کرد که در آن صورت چارچوب، فهرست شماره‌گذاری‌شدهٔ این قطعات است.

۶.۱ تعریف سرشماری

سرشماری از جامعهٔ متناهی، بررسیی است که همهٔ واحدهای جامعه را در بر می‌گیرد. در بسیاری از موارد، اجرای سرشماری در یک جامعهٔ متناهی کاری است شدنی. مثلاً در اداره‌ای که ۱۰۰ نفر کارمند دارد به‌سهولت می‌توان نظر مثبت یا منفی تمام آنها را نسبت به مادهٔ جدیدی که قرار است به قانون کار اضافه کنند پرسید. جمع‌آوری این نظرات یک سرشماری است. سرشماری را گاهی تمام‌شماری نیز می‌نامند.

وقتی حجم جامعه بزرگ باشد انجام سرشماری غالباً وقتگیر و پرهزینه است. بنابراین جز در موارد خیلی ضروری، برای بررسی یک مشخصهٔ جامعه از نمونه‌گیری استفاده می‌شود. در ایران هر ۱۰ سال یک بار سرشماری به‌عمل می‌آید و در طول ده سال بعد با نمونه‌گیریهای مقطعی تغییرات تدریجی را بررسی می‌کنند.

۷.۱ مزایای نمونه‌گیری

مزایای اصلی نمونه‌گیری در مقایسه با سرشماری به‌شرح زیرند:

الف. کاهش هزینه. چون حجم داده‌های نمونه‌ای نسبت به‌حجم جامعه کوچک است، محققاً هزینهٔ تهیهٔ آنها به‌مراتب کمتر از هزینهٔ سرشماری است. معمولاً نتایجی که از طریقهٔ نمونه‌گیری حاصل می‌شوند، آن‌قدر دقیق‌اند که می‌توان آنها را به‌عنوان نتایج سرشماری مورد استفاده قرار داد. در آمریکا معمولاً در بررسیهای نمونه‌ای غالباً از هر ۱۲۴۰ واحد یک واحد برای عضویت در نمونه انتخاب می‌شود، و مسلماً هزینهٔ چنین نمونه‌ای کمتر از هزینهٔ سرشماری از جامعهٔ اصلی است.

ب. افزایش سرعت کار. چون حجم نمونه بسیارکمتر از حجم جامعه در سرشماری است، گردآوری و تلخیص داده‌ها با سرعت بیشتر، یعنی با صرف وقت کمتر انجام می‌شود. این مزیّت، خصوصاً وقتی که کسب نتایج و تهیهٔ اطلاعات جنبهٔ فوریت دارند مزیتی مهم است.

ج. بالا بودن توان کار. در برخی از نمونه‌گیریها که وجود افراد متخصص و آموزش دیده و همچنین وجود وسایل اندازه‌گیری و انجام آزمونهای دقیق برای گردآوری داده‌ها ضروری است مسلماً به‌دلیل کمبود امکانات، انجام سرشماری عملاً غیرممکن است. مثلاً فرض کنید که بخواهند نسبت بیماران مبتلا به سل را در منطقهٔ باختران برآورد کنند. عملاً امتحان ب.ث.ژ و عکسبرداری از ششهای همهٔ افراد ممکن نیست زیرا در حالی که به‌افراد فنی برای آزمایش و عکسبرداری، و به‌دستگاههای متعدد نیاز است، تعداد دستگاههای عکسبرداری و تکنیسینهای این رشته محدود است. اما اگر نمونه‌گیری به‌عمل آید، چون تعداد کمی باید آزمایش شوند می‌توان از افراد متخصص و مجرب استفاده کرد که این استفاده، توان کاری بالاتری را موجب می‌شود.

د. بالا بودن میزان درستی کار. چون برای انجام نمونه‌گیری به‌دلیل کم بودن حجم کار نسبت به سرشماری، امکان آموزش افراد برای انجام مصاحبه‌ها و یا اندازه‌گیریها و غیره وجود دارد، میزان درستی نتیجهٔ کار این افراد در مقایسه با سرشماری بیشتر است.

ه. پرهیز از خراب کردن واحدهای جامعه. در بعضی از جامعه‌ها اصولاً امکان انجام سرشماری نیست زیرا ممکن است به از بین رفتن همهٔ واحدهای جامعه منجر شود. در چنین جامعه‌هایی ناگزیریم برای برآورد مشخصه‌ای از جامعه، صرفاً از نمونه‌گیری استفاده کنیم. مثلاً اگر هدف، بررسی نسبت لامپهای سالم یک بار مصرف فلاش عکاسی در یک محموله باشد نمی‌توان به‌عنوان سرشماری همهٔ لامپها را آزمایش کرد زیرا همهٔ لامپها از بین می‌روند. بدیهی است که با انجام نمونه‌گیری ضمن امکان تهیهٔ برآورد نسبت لامپهای سالم، قسمت عمدهٔ لامپهای محموله سالم می‌مانند.

۸.۱ مراحل اصلی در بررسی نمونه‌ای

به‌عنوان مقدمه‌ای بر نقشی که نظریهٔ نمونه‌گیری در یک بررسی دارد، توضیح مراحل انجام یک نمونه‌گیری ضروری است. کار نمونه‌گیری کاری پیچیده است. گرفتن نمونه‌ای از ۱۰۰۰۰ پروندهٔ شماره‌دار مثلاً کارکنان بانک ملی کاری ساده است، اما اگر هدف گرفتن نمونه‌ای از ساکنین یک ناحیه یا دهکده‌هایی با راههای صعب‌العبور و گاهی با زبانهای محلّی متفاوت باشد کار تهیهٔ نمونه ساده نیست، بخصوص که بسیاری از واحدهای نمونه ساکنینی هستند که افراد غریبه را نمی‌پذیرند و برای اظهار حقایق به آنها بدگمان‌اند. در هر حال مراحل اصلی یک بررسی نمونه‌ای به‌شرح زیرند:

۱. هدفهای بررسی. همیشه باید دربارهٔ هدفهای بررسی حکمی روشن و گویا در دست باشد. در غیر این صورت با افزایش حجم کار و جزئیات دیگر نمونه‌گیری تصمیمهایی گرفته می‌شوند که با اصل هدفها هماهنگی نخواهند داشت.

۲. جامعهٔ مورد نمونه‌گیری. جامعه‌ای را که می‌خواهیم از آن نمونه بگیریم باید دقیقاً تعریف کنیم. با این تعریف متغیر تحت بررسی مشخص می‌شود. البته تعریف دقیق جامعه همیشه امکان ندارد، گرچه گاهی بدون مشکل انجام می‌شود. مثلاً اگر هدف ما تعیین برآورد متوسط لامپهایی باشد که در یک روز به‌وسیلهٔ کارخانه‌ای ساخته می‌شوند، تعریف جامعه کاملاً روشن است. در حالی که نمونه‌گیری از جامعهٔ کشاورزی، مستلزم ارائه قاعده‌هایی است که باید مزرعه، مساحت مزرعه، و مرزهای آن را تعریف کنند. مسلماً این قاعده‌ها باید در عمل قابل پیاده کردن باشند. گاهی اوقات جامعه‌ای که از آن نمونه‌گیری می‌کنیم، دقیقاً همان جامعه‌ای نیست که می‌خواهیم دربارهٔ آن برآوردهایی تهیه کنیم. جامعه‌ای که می‌خواهیم برای مشخصه‌های آن برآوردهایی به‌دست آوریم **جامعهٔ هدف** نامیده می‌شود. ایدآل آن است که جامعه‌ای که از آن نمونه می‌گیریم همان جامعهٔ هدف باشد ولی همان‌طور که قبلاً هم متذکر شدیم به‌دلایل عملی و تغییر روزبه‌روز جامعه، جامعهٔ تحت

نمونه‌گیری محدودتر از جامعهٔ اصلی است. این محدودیت نیز گاهی به‌دلیل تسهیل کار به‌وجود می‌آید. خاطرنشان می‌کنیم که وقتی با چنین وضعی مواجه می‌شویم باید بررسی کنیم که آیا نتایج حاصل از نمونه، دربارهٔ جامعهٔ هدف هم صادق‌اند یا نه. این بررسی و قضاوت به منابع اطلاعاتی اضافی نیاز دارد. هر اطلاعی دربارهٔ طبیعت تفاوت بین جامعهٔ تحت نمونه‌گیری و جامعهٔ هدف، برای این بررسی و قضاوت مفید است.

۳. چارچوب. قبلاً توضیح دادیم که غالباً نمونه‌گیری در جامعهٔ هدف انجام نمی‌شود. تعریف چارچوب نمونه‌گیری را نیز عرضه کردیم. قبل از انتخاب نمونه باید فهرست واحدهای نمونه، یعنی چارچوب را مشخص کنیم، در واقع باید واحد نمونه‌گیری را تعریف کرد. کل این واحدها چارچوب را مشخص می‌کنند. واحدها نباید متداخل باشند. تعیین چارچوب از عمده‌ترین کارهای نمونه‌گیری است. اگر چارچوب دقیقاً تعریف نشود، گاهی شامل واحدهایی می‌شود که دوبار منظور شده‌اند. تهیهٔ چارچوب باید به‌وسیلهٔ افراد خبره، با توجه به موقعیتها صورت گیرد. در هر حال چارچوب باید دربردارندهٔ همهٔ واحدهای مورد نظر و فاقد واحدهای غیرلازم باشد.

۴. درجهٔ دقت مطلوب. نتایج یک بررسی نمونه‌ای همیشه با عدم حتمیت همراه است، زیرا اولاً بخشی از جامعه مورد اندازه‌گیری قرارگرفته است و ثانیاً اندازه‌گیریها همیشه با خطا همراه‌اند. بدیهی است با توجه به‌روش نمونه‌گیری، از وجود قانونهای احتمال حاکم بر نمونه برای کاستن میزان عدم حتمیت و تهیهٔ برآوردهای بازه‌ای استفاده می‌کنیم. میزان عدم دقت را می‌توان با افزایش حجم نمونه و با استفاده از وسایل دقیقتر اندازه‌گیری و گماردن افراد خبره برای انجام کار، تقلیل داد. مسلماً این کار با افزایش هزینه و با صرف وقت بیشتر انجام می‌شود، و این افزایش هزینه و صرف وقت بیشتر، نسبت مستقیم با میزان دقتی دارد که مایلیم برآوردهای مورد نظر داشته باشند. بنابراین مشخص کردن درجهٔ دقت مطلوب برآوردها گامی مهم در اجرای نمونه‌گیری است. این تشخیص میزان دقت، وظیفهٔ استفاده کنندگان از داده‌ها و برآوردهاست. متأسفانه اکثر مدیرانی که از نتایج نمونه‌گیری استفاده می‌کنند بصیرت آماری لازم را ندارند و بدین علت غالباً قادر به تعیین میزان درجهٔ دقت مطلوب نیستند.

۵. انتخاب روش نمونه‌گیری. با انجام مراحل بالا، به‌مرحله‌ای می‌رسیم که باید واحدهای نمونه را از بین واحدهای چارچوب انتخاب کنیم. برای انتخاب نمونه، طرحها و روشهای متعددی وجود دارند. برای هر چارچوبی با توجه به اندازهٔ واحدها و میزان تغییر آنها روشی خاص مناسب است. این روشها موضوع اصلی بحث این کتاب‌اند و در موقع مناسب، مناسبت هر روش را برای موقعیت مورد نظر بررسی خواهیم کرد. با توجه به‌دقت مورد نظر در محاسبهٔ برآوردها، باید حجم خاصی برای نمونه در نظر گرفت و با توجه به‌حجم لازم برای تأمین دقت مورد نظر، هزینهٔ لازم را پیش‌بینی کرد.

۶. روشهای اندازه‌گیری. در هر بررسی نمونه‌ای، برای اندازه‌گیری واحدهای نمونه، انتخاب و تهیهٔ ابزار اندازه‌گیری و مشخص کردن روشهای اندازه‌گیری از اهمیتی خاص برخوردار است. داده‌هایی که مربوط به وضعیت سلامتی یک فردند باید از آزمایشهای پزشکی به‌دست آیند و آزمایشهای پزشکی با ابزارهای خاص و با روشهای خاص انجام می‌شوند. البته ممکن است بررسی نمونه‌ای صرفاً به‌وسیلهٔ پرسشنامه و یا به‌وسیلهٔ مصاحبه با طرح سؤالاتی انجام شود که از پیش تعیین شده‌اند. مصاحبه ممکن است حضوری، یا با تلفن، و یا با ترکیبی از هر دو صورت گیرد. در هر حال پرسشنامه یا مصاحبه هم وسیله‌ای برای اندازه‌گیری هستند. بخش مهمی از کارهای اولیه نمونه‌گیری تهیهٔ وسایل اندازه‌گیری، تنظیم فرمهای ثبت داده‌هاست که باید پس از اندازه‌گیری مشخصه‌ها پاسخها در آنها منعکس شوند. تنظیم فرمها باید به‌صورتی باشد که انتقال مستقیم داده‌ها به کامپیوتر را میسر سازد. در واقع برای تنظیم خوب فرمهای ثبت داده‌ها لازم است که ساختار جداول نهایی خلاصه‌ها را که برای استخراج نتایج ضروری‌اند از ابتدا در نظر داشته باشیم.

۷. آموزش آمارگیران. در بررسیهای جامع نمونه‌ای، اغلب با مسائل خاص حرفه‌ای روبه‌رو هستیم. لذا آمارگیران باید قبلاً دربارهٔ هدف نمونه‌گیری، واحد نمونه‌گیری، روشهای اندازه‌گیری، کار با دستگاههای اندازه‌گیری و پرسش سؤالها و یا انجام مصاحبه‌ها و نحوهٔ گردآوری داده‌ها و سایر خط‌مشی‌ها آموزش ببینند. بعد از آموزش، و پس از نمونه‌گیری مقدماتی که در زیر شرح می‌دهیم مجدداً جلسات توجیهی برای آمارگیران ضروری است. نتیجهٔ این کار ایجاد یک درک یکسان از سؤالها و یکنواختی در عمل آمارگیری است.

۸. پیش‌آزمون. تجربه نشان داده است که وقتی دقت مورد نظر را از قبل مشخص و روش نمونه‌گیری و حجم نمونه را معین کردیم، لازم است که قبل از انجام نمونه‌گیری، برای بررسی کارایی پرسشنامه و روش کار و مشکلات اجرایی، عمل نمونه‌گیری را در مقیاس کوچک انجام دهیم. این نمونه‌گیری با حجم کوچک مقدماتی نتایجی به‌دست می‌دهد که براساس آنها می‌توان پرسشنامه‌ها را اصلاح کرد، به آمارگیران آموزش تکمیلی داد، در مورد زمان لازم برای نمونه‌گیری اصلی برآوردی به‌دست آورد، هزینهٔ عملی نمونه‌گیری را بازبینی نمود، و در نتیجه از بروز اشکالات عمده در نمونه‌گیری اصلی که حجمی زیاد دارد جلوگیری کرد. آمارگیران شخصاً در این نمونه‌گیری مقدماتی مشکلات کار را می‌بینند و نوع عکس‌العملها را فرا می‌گیرند. این فرایند نمونه‌گیری با حجم کم را نمونه‌گیری مقدماتی و یا پیش آزمون می‌نامند که معمولاً با هزینهٔ کم مانع از به‌هدر رفتن هزینهٔ زیاد نمونه‌گیری اصلی می‌شود.

۹. گردآوری داده‌ها. برای گردآوری داده‌ها معمولاً از پرسشنامه استفاده می‌کنند. تهیهٔ متن پرسشها از نظر تطبیق با اهداف، ایجاد انگیزه برای پاسخ، جلوگیری از ایجاد انحراف در پاسخها و بسیاری نکات ظریف دیگر کاری فنّی است و باید پرسشنامه به‌وسیلهٔ افراد خبره با توجه به روانشناسی پاسخگویان

تهیه شود. داده‌های حاصل باید از نظر اینکه دقیقاً در راستای اهداف بررسی باشند کنترل شوند. نباید داده‌ای اساسی از قلم بیفتد. گاهی در پرسشنامه‌ها یا مصاحبه‌ها سؤالاتی مطرح می‌شوند که اصولاً در تحلیلهای بعدی مورد استفاده قرار نمی‌گیرند. یک پرسشنامهٔ طویل از کیفیت پاسخها می‌کاهد.

۱۰. **پردازش داده‌ها.** اولین گام در این مقطع، آماده کردن پرسشنامه‌های تکمیل‌شده و یا هر نوع دادهٔ دیگر برای انتقال به ماشین است. در این مرحله خطاهایی که ثبت شده‌اند باید اصلاح شوند و پرسشنامه‌هایی که در آنها، به‌وضوح برخی اقلام اشتباهاً ثبت شده‌اند حذف و با مراجعهٔ مجدد اصلاح شوند. دربارهٔ چگونگی رفتار با سؤالاتی که پاسخ‌دهنده به آنها جواب نداده است و یا در موقع انتقال آنها به ماشین حذف شده‌اند باید تصمیم‌گیری کرد. بعد از ورود داده‌ها به ماشین باید براساس روش نمونه‌گیری و آماره‌های مورد نظر و فرمولهای محاسباتی متناظر، محاسباتی که برای تهیهٔ برآوردها لازم‌اند انجام شوند. در این مقطع از زبانهای مختلف کامپیوتری برحسب مورد می‌توان استفاده کرد. نرم‌افزارهای متعدد قابل اطمینان به‌صورت بسته‌های کامپیوتری آماری وجود دارند. در ادامهٔ استخراج نتایج، محاسبهٔ میزان خطا در برآوردها ضروری است. چون روشهای نمونه‌گیری مورد استفاده، روشهای احتمالاتی هستند، همیشه امکان بیان احکامی احتمالاتی دربارهٔ بازه‌های اطمینان پارامترهای برآوردشده وجود دارند و چگونگی دستیابی به این احکام معمولاً در بسته‌های کامپیوتری آماری منظور شده‌اند.

۱۱. **اطلاعات حاصل برای بررسیهای آینده.** هر نمونه‌ای که از جامعه گرفته می‌شود علاوه بر نتایج استنباطی آن، راهنمایی بالقوه، برای بهتر کردن فرایند نمونه‌گیریهای بعدی است. طبعاً از این نمونه می‌توان دربارهٔ میانگینها، انحراف معیارها، رفتار و تغییرپذیری اندازه‌ها و هزینهٔ گردآوری داده‌ها اطلاعاتی برای استفاده در طرّاحی نمونه‌گیریهای آینده به‌دست آورد. در هر نمونه‌گیری کارها دقیقاً همان‌گونه که طرّاحی شده‌اند انجام نمی‌شوند، با بررسی فرایند نمونه‌گیری انجام‌شده، می‌توان خطاهایی را که موجب این مشکل شده‌اند شناسایی و از تکرار آن در نمونه‌گیریهای بعدی جلوگیری کرد.

ما از جزئیات همهٔ این مراحل بحث نمی‌کنیم، اما خاطرنشان می‌کنیم که ضعف کار در هر کدام از این مراحل می‌تواند نتیجهٔ مراحل دیگر را تحت تأثیر قرار دهد. همان‌طور که متذکر شدیم موضوع اصلی بحث این کتاب شرح روشهای مختلف نمونه‌گیری و ارائهٔ بعضی استنباطهایی است که، از روی نمونهٔ حاصل از این روشها، دربارهٔ جامعه به‌دست می‌آیند. قبل از آغاز این کار برخی از تعریفها و مفاهیمی را که در آمار مقدماتی و آمار ریاضی مطرح می‌شوند و برای بسط روشهای نمونه‌گیری ضروری‌اند یادآوری می‌کنیم.

۹.۱ مروری بر برخی مفاهیم آماری

در نظریهٔ نمونه‌گیری، هدف، برآورد پارامترهای جامعه از روی داده‌های نمونه‌ای، و بیان حکمی احتمالاتی دربارهٔ دقت برآورد است. در کتاب حاضر، مفاهیم متغیر تصادفی، بردار تصادفی، توزیع

یک‌متغیره و چندمتغیره و توزیع شرطی را دانسته فرض می‌کنیم و برخی ویژگیهای این مفاهیم و مطالب وابسته به آنها راکه در توضیح روشهای مختلف نمونه‌گیری مورد نیازند به‌اختصار مرور می‌کنیم.

این مرور، برای سادگی کار، عمدتاً در حالت گسسته انجام می‌شود. مطالبی که یادآوری آنها ضروری است به‌شرح زیرند:

۱. **امید ریاضی.** اگر X متغیری تصادفی باشد که مقادیر $x_i(i = ۱,\ ۲, \ldots, k)$ را با احتمالهای متناظر $p_i(i = ۱,\ ۲, \ldots, k)$، با شرایط $۰ \leq p_i \leq ۱$، $\sum_i p_i = ۱$ اختیار کند، آن‌گاه امید ریاضی X و یا ساده‌تر، امید X به‌صورت

$$E(X) = \sum_i p_i x_i = \overline{X} \qquad (۱.۱)$$

تعریف می‌شود. به‌ازای مقدارهای ثابت a و b، داریم

$$E(aX + b) = aE(X) + b \qquad (۲.۱)$$

اگر $G(X)$ تابعی از متغیر تصادفی X باشد،

$$E[G(X)] = \sum_i p_i\, G(x_i) \qquad (۳.۱)$$

به‌خصوص اگر $G(x)$ تابعی محدب باشد آن‌گاه برای متغیر تصادفی X، نابرابری معروف به ینسن[1] به‌صورت زیر برقرار است

$$E(G(X)] \geq G(E(X)) \qquad (۳.۱)'$$

فرض می‌کنیم Y متغیر تصادفی دیگری باشد که مقادیر $y_j(j = ۱,\ ۲, \ldots, L)$ را با احتمالهای متناظر $p_j(j = ۱,\ ۲, \ldots, L)$، با شرایط $۰ \leq p_j \leq ۱$، $\sum_j p_j = ۱$ اختیار کند. به‌علاوه فرض می‌کنیم توزیع توأم زوج $(X,\ Y)$ با رابطهٔ

$$p_{ij} = \Pr(X = x_i, Y = y_j) \quad , \quad \sum_i \sum_j p_{ij} = ۱ \qquad (۴.۱)$$

داده شود. در این صورت

$$p_i = \Pr(X = x_i) = \sum_j p_{ij} \quad , \quad p_j = \Pr(Y = y_j) = \sum_i p_{ij} \qquad (۵.۱)$$

1. Jensen

و ثابت می‌شود که اگر $Z = X + Y$. آن‌گاه

$$E(Z) = E(X+Y) = E(X) + E(Y) \tag{۶.۱}$$

این رابطه به استقلال یا عدم استقلال X و Y بستگی ندارد، و به‌چند متغیر تصادفی نیز تعمیم‌پذیر است. یعنی در حالت کلی اگر X_1، X_2، $\ldots$ ، X_n متغیرهایی تصادفی باشند

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \tag{۷.۱}$$

اگر X و Y مستقل باشند، داریم

$$\Pr(X = x_i,\ Y = y_j) = \Pr(X = x_i) \cdot \Pr(Y = y_j) \tag{۸.۱}$$

یا

$$p_{ij} = p_i \cdot p_j \tag{۹.۱}$$

در نتیجه داریم

$$E(XY) = E(X) \cdot E(Y) \tag{۱۰.۱}$$

زیرا

$$\begin{aligned} E(XY) &= \sum_i \sum_j p_{ij} \cdot x_i y_j = \sum_i \sum_j p_i p_j x_i y_j \\ &= \sum_i p_i x_i \cdot \sum_j p_j Y_j = E(X)E(Y) \end{aligned}$$

به‌طور کلی اگر $f(X)$ و $G(Y)$ دو تابع از متغیرهای مستقل X و Y باشند، داریم

$$E[f(X) \cdot G(Y)] = E[f(X)] \cdot E[(G(Y)] \tag{۱۱.۱}$$

در صورتی که X و Y مستقل نباشند

$$E(XY) = E[XE(Y|X)] = E[YE(X|Y)] \tag{۱۲.۱}$$

زیرا مثلاً برای صورت اول

$$E(XY) = \sum_i \sum_i p_{ij} x_i y_j = \sum_i x_i \sum_j p_{ij} y_j$$

اما با توجه

$$\Pr(Y = y_j | X = x_i) = \frac{p_{ij}}{p_i}$$

نتیجه می‌شود که

$$\begin{aligned} E(XY) &= \sum_i x_i \sum_j p_i \Pr(Y = y_j | X = x_i) y_j \\ &= \sum_i p_i x_i \sum_j \Pr(Y = y_j | X = x_i) y_j \\ &= \sum_i p_i x_i E(Y|X) = \sum_i p_i [x_i E(Y|X)] \\ &= E[X E(Y|X)] \end{aligned}$$

اگر X و Y مستقل باشند $E(Y|X) = E(Y)$ و رابطهٔ (۱۰.۱) به‌دست می‌آید.

۲. واریانس. اگر X متغیری تصادفی و $\overline{X} = E(X)$ باشد، آنگاه واریانس X به‌صورت

$$V(X) = \sum_i p_i (x_i - \overline{X})^2 = E[(X - \overline{X})^2] = E(X^2) - \overline{X}^2 \quad (۱۳.۱)$$

تعریف می‌شود. اگر $V(X)$ کوچک باشد عمدتاً هر جمله از مجموع $\sum_i p_i (x_i - \overline{X})^2$ کوچک است، بدین معنا که لازم است مقداری از X که برای آن $|x_i - \overline{X}|$ بزرگ است با احتمال کوچک p_i همراه باشد. بنابراین احتمال وجود انحرافهای بزرگ از میانگین کم خواهد بود.

برای متغیر تصادفی $Y = X + b$

$$V(Y) = V(X + b) = V(X)$$

و برای متغیر تصادفی $Y = aX + b$

$$V(Y) = V(aX + b) = a^2 V(X) \quad (۱۴.۱)$$

ریشهٔ مثبت $V(X)$، انحراف معیار X است که آن را با نماد $\sigma(X)$ نشان می‌دهیم. اگر X_1، X_2، ...، X_n متغیرهایی تصادفی و دوبه‌دو مستقل باشند

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) \quad (۱۵.۱)$$

۳. کوواریانس و ضریب همبستگی. مفهوم دیگری که در روشهای نمونه‌گیری کاربرد زیاد دارد کوواریانس بین متغیرهای تصادفی است. کوواریانس بین دو متغیر تصادفی X و Y به‌صورت زیر تعریف می‌شود

$$\begin{aligned}\operatorname{cov}(X,\ Y) &= \sum_i \sum_j p_{ij}(x_i - \overline{X})(y_j - \overline{Y}) = E[(X - \overline{X})(Y - \overline{Y})] \\ &= E(XY) - E(X)E(Y) \qquad (۱۶.۱)\end{aligned}$$

اگر X و Y مستقل باشند، بنابر (۱۰.۱)، $E(XY) = E(X)E(Y)$ و درنتیجه $\operatorname{cov}(X,\ Y) = ۰$.
بدیهی است که

$$\operatorname{cov}(X,\ X) = \mathrm{V}(X)$$

اگر کوواریانس را بر حاصلضرب انحراف معیارهای $\sigma(X)$ و $\sigma(Y)$ تقسیم کنیم، ρ، ضریب همبستگی تعریف می‌شود. یعنی

$$\rho(X,\ Y) = \frac{\operatorname{cov}(X,\ Y)}{\sigma(X)\sigma(Y)} \qquad (۱۷.۱)$$

ρ در بازهٔ $(۱ ,۱-)$ تغییر می‌کند. اگر X و Y مستقل باشند $\rho = ۰$، ولی اگر داشته باشیم $\rho = ۰$، نتیجه می‌شود $\operatorname{cov}(X,\ Y) = ۰$، که نمی‌توان از آن، در حالت کلی استقلال X و Y را نتیجه گرفت. وقتی $\rho = ۰$، متغیرهای X و Y را ناهمبسته گویند. اگر توزیع توأم $(X,\ Y)$ نرمال دومتغیره باشد شرط $\rho = ۰$، استقلال X و Y را نتیجه می‌دهد. با توجه به بازهٔ تغییر ρ، همیشه $|\rho(X,\ Y)| \leq ۱$. با توجه به‌مفهوم کوواریانس، می‌توانیم واریانس مجموع چند متغیر را در حالت کلی بررسی کنیم.

اگر $Y_i(i = ۱,\ ۲, \ldots ,n)$، n متغیر تصادفی و $a_i(i = ۱,\ ۲, \ldots ,n)$، n مقدار ثابت باشند

$$\mathrm{V}\left(\sum_i a_i X_i\right) = \sum_i \sum_j a_i a_j \operatorname{cov}(X_i,\ X_j) \qquad (۱۸.۱)$$

با توجه به اینکه وقتی i و j مقادیر برابر اختیار می‌کنند $\operatorname{cov}(X_i,\ X_i) = \mathrm{V}(X_i)$، می‌توان (۱۸.۱) را به‌صورت زیر هم نوشت

$$\mathrm{V}\left(\sum_i a_i X_i\right) + \sum_i a_i^2 \mathrm{V}(X_i) + ۲\sum_{\substack{i \\ i<j}} \sum_j a_i a_j \operatorname{cov}(X_i,\ X_j) \qquad (۱۹.۱)$$

یا

$$\mathrm{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^{۲}\mathrm{V}(X_i) + ۲\sum_i \sum_{\substack{j \\ j>i}} a_i a_j \rho_{ij}\sigma(X_i)\sigma(X_j) \qquad (۲۰.۱)$$

که در آن ρ_{ij} ضریب همبستگی بین دو متغیر تصادفی X_i و X_j است. اگر X_iها دوبه‌دو ناهمبسته هم باشند، یعنی $\rho_{ij} = ۰$، آن‌گاه

$$\mathrm{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^{۲}\mathrm{V}(X_i)$$

که با حالتی که متغیرها مستقل باشند مطابقت می‌کند. رابطهٔ بالا مبین واریانس مجموع متغیرهای تصادفی است، اگر X و Y دو متغیر تصادفی مستقل باشند می‌توان عبارتی ساده برای واریانس حاصلضرب آنها یافت. فرض کنید $E(X) = \overline{X}$، $E(Y) = \overline{Y}$. قرار می‌دهیم

$$\frac{X - \overline{X}}{\overline{X}} = \delta x \quad , \quad \frac{Y - \overline{Y}}{\overline{Y}} = \delta y$$

چون X و Y مستقل فرض شده‌اند $E(XY) = \overline{X}\,\overline{Y}$. از طرفی با توجه به دو قرارداد بالا

$$X = \overline{X}(۱ + \delta x) \quad , \quad Y = \overline{Y}(۱ + \delta y)$$

پس

$$\begin{aligned}
\mathrm{V}(XY) &= E[(XY - \overline{X}\,\overline{Y})^{۲}] = E[\{\overline{X}\,\overline{Y}(۱+\delta x)(۱+\delta y) - \overline{X}\,\overline{Y}\}^{۲}] \\
&= (\overline{X}\,\overline{Y})^{۲} E[(\delta x + \delta y + \delta x \cdot \delta y)^{۲}] \\
&= (\overline{X}\,\overline{Y})^{۲}\Bigg\{E\left[\left(\frac{X-\overline{X}}{\overline{X}}\right)^{۲}\right] + E\left[\left(\frac{Y-\overline{Y}}{\overline{Y}}\right)^{۲}\right] \\
&\qquad + E\left[\left(\frac{X-\overline{X}}{\overline{X}} \cdot \frac{Y-\overline{Y}}{\overline{Y}}\right)^{۲}\right]\Bigg\}
\end{aligned}$$

در آخرین عبارت، بقیهٔ جمله‌ها به‌دلیل استقلال X و Y، صفر می‌شوند. پس

$$\mathrm{V}(XY) = (\overline{X}\,\overline{Y})^{۲}\left[\frac{\mathrm{V}(X)}{\overline{X}^{۲}} + \frac{\mathrm{V}(Y)}{\overline{Y}^{۲}} + \frac{\mathrm{V}(X)\cdot\mathrm{V}(Y)}{\overline{X}^{۲}\,\overline{Y}^{۲}}\right]$$

که به نتیجهٔ مهم زیر می‌رسیم

$$V(XY) = \overline{Y}^{2} V(X) + \overline{X}^{2} V(Y) + V(X)V(Y) \qquad (۲۱.۱)$$

اگر دو تابع خطی از مجموعهٔ متغیرهای تصادفی $X_i(i=1, 2, \dots, n)$ و $Y_i(i=1, 2, \dots, m)$ به‌صورت زیر داشته باشیم

$$X = \sum_{i=1}^{n} a_i X_i \quad , \quad Y = \sum_{j=1}^{m} b_j Y_j$$

آن‌گاه

$$\operatorname{cov}(X, Y) = \sum_i \sum_j a_i b_j \operatorname{cov}(X_i, Y_j) \qquad (۲۲.۱)$$

۴. میانگین شرطی. اگر X و Y یک زوج متغیر تصادفی باشند،

$$E(X) = E_Y[E_X(X|Y)] \qquad (۲۳.۱)$$

که به امید ریاضی دوگانه موسوم است. نماد $E_X(X|Y)$، به‌معنای آن است که Y مقداری تثبیت‌شده اختیار می‌کند و امید ریاضی نسبت به X محاسبه می‌شود که نتیجه تابعی از Y است. وقتی امید این تابع با توجه به‌توزیع احتمال Y حساب شود نتیجهٔ حاصل همان امید X است. به‌عبارت دیگر، متوسط‌گیری از میانگینهای شرطی نسبت به‌متغیر شرطی‌کننده، میانگین غیرشرطی را به‌دست می‌دهد. به‌همین ترتیب داریم

$$V(X) = E[V(X|Y)] + V[E(X|Y)] \qquad (۲۴.۱)$$

که دارای این مفهوم است که واریانس X عبارت از میانگین واریانس شرطی به‌علاوهٔ واریانس میانگین شرطی آن است. در روشهای نمونه‌گیری، به‌ویژه نمونه‌گیری خوشه‌ای، برای اثبات برخی ویژگیها از رابطه‌های (۲۳.۱) و (۲۴.۱) استفاده می‌شود.

۵. ضریب همبستگی برای زوج متغیرهای تصادفی. وقتی از زوج متغیر تصادفی گسستهٔ (X, Y) نمونهٔ تصادفی (X_1, Y_1)، (X_2, Y_2)، . . . ، (X_n, Y_n) را می‌گیریم، چون معمولاً X و Y دو مشخصه از واحدهای جامعه‌اند در حالت کلی به‌هم وابسته‌اند. قرار می‌دهیم

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{N} (X_i - \overline{X}_N)(Y_i - \overline{Y}_N) \qquad (۲۵.۱)$$

که در آن N حجم جامعه و $\overline{X}_N$ و $\overline{Y}_N$ به‌ترتیب میانگینهای دو مشخصهٔ X و Y در جامعه‌اند. می‌دانیم ضریب همبستگی دو مشخصه

$$\rho(X, Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{۱}{N}\sum_{i=۱}^{N}(X_i - \overline{X}_N)(Y_i - \overline{Y}_N)}{\left[\frac{۱}{N}\sum_{i=۱}^{N}(X_i - \overline{X}_N)^۲ \cdot \frac{۱}{N}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲\right]^{\frac{۱}{۲}}}$$

اگر طبق معمول قرار دهیم

$$S_X^۲ = \frac{۱}{N-۱}\sum_{i=۱}^{N}(X_i - \overline{X}_N)^۲ \quad , \quad S_Y^۲ = \frac{۱}{N_۱}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲ \qquad (۲۶.۱)$$

با منظور کردن (۱.۲۶) و (۱.۲۵) در ضریب همبستگی، نتیجه می‌شود که

$$\rho(X, Y) = \frac{S_{XY}}{S_X S_Y} \qquad (۲۷.۱)$$

مثال ۳.۱ نمرات امتحان دروس استنباط آماری و احتمال پیشرفتهٔ ۶ نفر دانشجویان دورهٔ کارشناسی ارشد یک بخش آمار به‌صورت زیر است

$$(۱۸, ۱۲), (۱۶, ۱۴)\ (۱۷, ۱۳), (۱۴, ۱۶), (۱۴, ۱۴), (۱۷, ۱۵)$$

اگر متغیر تصادفی X معرّف نمرهٔ استنباط آماری و متغیر تصادفی Y معرف نمرهٔ احتمال پیشرفته باشد زوج (X, Y) متغیر تصادفی زوجی است و حجم جامعه برابر با ۶ است. برای محاسبهٔ ضریب همبستگی بین نمرات دو درس در این کلاس به‌ترتیب $S_X^۲$، $S_Y^۲$ و S_{XY} را حساب می‌کنیم.

$$S_X^۲ = \frac{۱}{۵}[(۱۸-۱۶)^۲ + (۱۶-۱۶)^۲ + \cdots + (۱۷-۱۶)^۲] = \frac{۱۴}{۵}$$
$$S_Y^۲ = \frac{۱}{۵}[(۱۲-۱۴)^۲ + (۱۴-۱۴)^۲ + \cdots + (۱۵-۱۴)^۲] = ۲$$
$$S_{XY} = \frac{۱}{۵}[(۱۸-۱۶)(۱۲-۱۴) + \cdots + (۱۷-۱۶)(۱۵-۱۴)]$$
$$= \frac{۱}{۵}(-۴-۱-۴+۱) = \frac{-۸}{۵}$$

بنابراین

$$\rho(X,Y) = \frac{-\frac{۸}{۵}}{\sqrt{\frac{۱۴}{۵}} \cdot \sqrt{۲}} \# -۰{,}۶۷۶$$

▲

۶. برآورد نقطه‌ای. اگر پارامتر یا مشخصه‌ای از جامعهٔ تحت بررسی مجهول باشد باید آن را از روی داده‌های نمونه‌ای برآورد کرد. ابتدا نمونهٔ تصادفی $X_i(i = ۱, \dots, n)$ را در نظر می‌گیرند و به‌کمک روشهایی که در بحث برآورد به‌کار می‌روند فرمولی برحسب X_iهای تصادفی برای برآورد پارامتر مجهول به‌دست می‌آورند. این فرمول را که تابعی از مشاهدات تصادفی است آماره می‌نامند. از قرار دادن یافته‌های X_iها در این آماره مقدار پارامتر مجهول به‌ازای داده‌های نمونه‌ای برآورد می‌شود. این آماره را برآوردکننده می‌نامیم. درواقع برآوردکننده، متغیری تصادفی است که اگر مقادیر نمونه‌ای را در آن قرار دهیم، رخدادی از آماره را که متناظر با نمونهٔ موجود است ارائه می‌دهد. این مقدار عددی آماره را برآورد پارامتر یا مشخصهٔ مجهول می‌گوییم. برآوردکننده‌ها به‌روشهایی مختلف تهیه می‌شوند که موضوع بحث آمار ریاضی هستند. این روشها عمدتاً روش درستنمایی ماکسیمم، روش کمترین مربعات، و روش گشتاورها و روش بیزی هستند. برآوردکننده‌ای از نظر آماری خوب است که در دو ملاک نااریبی، و سازگاری، صادق بوده و کارایی آن نسبت به برآوردکننده‌های دیگر بیشتر باشد. دربارهٔ این ویژگیها، توضیحی شهودی و نه‌چندان دقیق، در حدّی که مورد استفاده بحثهای این کتاب است، عرضه می‌شود.

الف) نااریبی. اگر پارامتر مجهول جامعه θ باشد و برای آن به‌طریقی برآوردکنندهٔ $\hat{\theta}$ را به‌دست آورده باشیم، در حالت کلی $\hat{\theta}$ متغیری تصادفی است. اگر میانگین $\hat{\theta}$ برابر با θ باشد آن را برآوردکننده‌ای نااریب می‌گویند و مقدار آن به‌ازای یافته‌های نمونه را برآورد نااریب θ می‌خوانند. واضح است که برای نمونه‌های مختلف به‌حجم n مقادیر مختلفی برای آمارهٔ $\hat{\theta}$ به‌دست می‌آید که همهٔ این مقادیر حول مقدار واقعی و مجهول θ واقع‌اند به‌گونه‌ای که

$$E(\hat{\theta}) = \theta$$

اگر میانگین $\hat{\theta}$ برابر با θ نباشد برآوردکننده را اریب می‌گویند. در این حالت

$$E(\hat{\theta}) = \theta + a$$

a را اریبی این برآوردکننده می‌نامند.

ب) سازگاری. اگر به‌کمک نمونه‌ای به‌حجم n، برآوردکنندهٔ $\hat{\theta}$ را برای پارامتر θی جامعه به‌دست آوریم انتظار نداریم که به‌ازای یافته‌های نمونه، برآورد θ دقیقاً برابر با مقدار واقعی θ باشد. اگر n را تدریجاً زیاد کنیم، یعنی برای جامعهٔ نامتناهی $n \longrightarrow \infty$ و برای جامعهٔ متناهی به‌حجم N، اجازه دهیم که $n \longrightarrow N$، ممکن است با بزرگ شدن حجم نمونه، مقدار $\hat{\theta}$ به مقدار واقعی θ میل کند. اگر چنین باشد، برآوردکننده را سازگار می‌گویند. اگر بخواهیم این ویژگی را به‌صورت احتمالاتی بیان کنیم، باید به‌ازای هر ε بینهایت کوچک مثبت، داشته باشیم

$$\lim_{n \longrightarrow \infty} \Pr(|\hat{\theta} - \theta| < \varepsilon) = ۱$$

اگر جامعه متناهی و به‌حجم N باشد رابطهٔ بالا باید به‌ازای $n \longrightarrow N$ برقرار شود.

مثال ۴.۱ اگر میانگین وزن $N = ۲۰۰۰$ دانشجو $\overline{Y}_N = ۷۰$ کیلوگرم باشد، و اگر حجم نمونه را $n = ۱۹۹۸$ بگیریم و میانگین نمونه را با $\overline{Y}_n$ نشان دهیم محققاً $\overline{Y}_n$ به $\overline{Y}_N = ۷۰$ نزدیک است، به‌ویژه اگر $n = ۱۹۹۹$ باشد تفاوت $\overline{Y}_n$ و $\overline{Y}_N$ ناچیز خواهد بود. یعنی $\overline{Y}_n$ وقتی $n \longrightarrow N$ برآوردکننده‌ای سازگار برای $\overline{Y}_N$ است.

حال اگر فرض کنیم میانهٔ جامعهٔ مزبور $\tilde{Y}_N = ۷۲$ باشد و نمونه‌ای به‌حجم $n = ۱۹۹۸$ از این جامعه انتخاب کنیم و میانهٔ آن یعنی $\tilde{Y}_n$ را برآوردکنندهٔ میانگین جامعه بگیریم. طبیعی است که $\tilde{Y}_n$ به میانهٔ جامعه یعنی به ۷۲ نزدیک است و نه به ۷۰ که میانگین جامعه است. لذا اگر $\tilde{Y}_n$ را برآوردکنندهٔ $\overline{Y}_N$ بگیریم این برآوردکننده برای $\overline{Y}_N$ سازگار نیست. اگر جامعه دارای توزیع نرمال باشد، چون در جامعهٔ نرمال، میانگین و میانه برابرند، میانهٔ نمونه هم برآوردکننده‌ای سازگار برای $\overline{Y}_N$ است.▲

ج) کارایی. اگر برای پارامتر مجهول θی جامعه، از روی یک نمونه، دو برآوردکنندهٔ $\hat{\theta}_۱$ و $\hat{\theta}_۲$ را داشته باشیم و

$$\mathrm{V}(\hat{\theta}_۱) < \mathrm{V}(\hat{\theta}_۲)$$

آن‌گاه ضریب کارایی $\hat{\theta}_۲$ نسبت به $\hat{\theta}$ را چنین تعریف می‌کنیم

$$e(\hat{\theta}_۲,\ \hat{\theta}_۱) = \frac{\mathrm{V}(\hat{\theta}_۱)}{\mathrm{V}(\hat{\theta}_۲)} \qquad ۰ \leq e \leq ۱ \tag{۲۸.۱}$$

به‌عبارت دیگر برآوردکننده‌ای که واریانس کمتری دارد کارایی بیشتری خواهد داشت. در بسیاری از موارد محاسبهٔ e و تعیین اینکه مقدار آن کوچکتر از یک است یا نه، به‌علت پیچیدگی صورتهای دو واریانس کار مشکلی است. در چنین مواردی معمولاً از تفاضل دو واریانس استفاده می‌کنیم. اگر $\mathrm{V}(\hat{\theta}_۱) - \mathrm{V}(\hat{\theta}_۲)$ مثبت باشد کارایی $\hat{\theta}_۲$ بیشتر است و اگر منفی باشد کارایی $\hat{\theta}_۱$ بیشتر است، اگر این تفاضل صفر باشد کارایی دو برآوردکننده یکسان است. در فصول آینده روشهای مختلف نمونه‌گیری در جامعه‌های متناهی تعریف و اجرا می‌شوند. مثال زیر گرچه مربوط به جامعهٔ نامتناهی است و نمونه‌گیری از آن به‌گونه‌ای است که در آمار ریاضی دیده‌اید ولی برای مقایسهٔ کارایی دو برآوردکننده مثالی گویاست.

مثال ۵.۱ اگر توزیع جامعهٔ تحت بررسی نرمال باشد، میانگین جامعه را می‌توان به‌گونه‌ای که در آمار ریاضی دیده‌اید با میانگین نمونه‌ای تصادفی به‌حجم n و یا با میانهٔ آن برآورد کرد. اگر میانگین نمونه و میانهٔ نمونه به‌ترتیب $\overline{Y}_n$ و $\tilde{Y}_n$ باشند هر دو برآوردکننده‌ای سازگار برای میانگین جامعه‌اند. در آمار ریاضی دیده‌اید که اگر $\sigma^۲$ واریانس جامعه باشد آن‌گاه

$$\mathrm{V}(\overline{Y}_n) = \frac{\sigma^۲}{n} \qquad , \qquad \mathrm{V}(\tilde{Y}_n) = \frac{\pi}{۲} \cdot \frac{\sigma^۲}{n}$$

لذا برای نمونه‌ای به‌حجم n

$$\frac{V(\overline{Y}_n)}{V(\tilde{Y}_n)} = \frac{\sigma^۲/n}{\pi\sigma^۲/۲n} = \frac{۲}{\pi} \# ۰{,}۶۴$$

یعنی کارایی $\tilde{Y}_n$ تقریباً برابر با ۶۴ر۰ کارایی $\overline{Y}_n$ است. ▲

در آمار ریاضی دیده‌اید که به‌طور کلی $\overline{Y}_n$ برای میانگین هر جامعه، برآوردکننده‌ای نااریب، سازگار و در اکثر موارد با واریانس مینیمم است. به‌عبارت دیگر اگر جامعه توزیع نرمال داشته باشد $\overline{Y}_n$ برآوردکنندهٔ با واریانس مینیمم میانگین جامعه است. برای توزیعهای دیگر در اکثر موارد چنین است ولی مواردی هم موجودند که $\overline{Y}_n$ برآوردکنندهٔ با واریانس مینیمم میانگین جامعه نیست. مثلاً اگر نمونه‌ای تصادفی به‌حجم ۳ از جامعه‌ای با توزیع یکنواخت

$$f(y) = \begin{cases} ۱ & \theta - \frac{۱}{۲} < y < \theta + \frac{۱}{۲} \\ ۰ & \text{سایر جاها} \end{cases}$$

بگیریم به‌سادگی می‌توان دید که θ برابر با میانگین جامعه است. بنابراین $\overline{Y}_n$ که از نمونه به‌حجم ۳ به‌دست می‌آید برآوردکنندهٔ نااریب θ است. از طرفی اگر $\frac{Y_{(۱)}+Y_{(۳)}}{۲}$ را به‌عنوان برآوردکنندهٔ θ بگیریم که در آن $Y_{(۱)}$ و $Y_{(۳)}$ به‌ترتیب کوچکترین و بزرگترین واحد نمونهٔ به‌حجم ۳ هستند به‌سهولت می‌توان تحقیق کرد که این برآوردکننده برای θ نااریب است. با به‌دست آوردن توزیع این برآوردکنندهٔ جدید و محاسبهٔ واریانس آن، خواهید دید که واریانسی کمتر از واریانس $\overline{Y}_n$ دارد، یعنی $\overline{Y}_n$، برآوردکنندهٔ θ با واریانس مینیمم نیست.

۷. برآورد بازه‌ای. پس از تعیین یک برآورد نقطه‌ای خوب برای یک پارامتر مجهول جامعه، نمی‌توان انتظار داشت که این برآورد دقیقاً با مقدار واقعی پارامتر برابر باشد. اگر مقدار واقعی پارامتر θ، و برآوردکنندهٔ خوب آن $\hat{\theta}$ باشد، رخدادهای $\hat{\theta}$ به‌ازای نمونه‌های مختلف هم حجم، حول θ پراکنده خواهند شد. هرچه دامنهٔ این پراکندگی یا انحراف معیار $\hat{\theta}$ کمتر باشد برآوردکنندهٔ $\hat{\theta}$ برآوردکنندهٔ بهتری برای θ است. معمولاً سعی می‌کنیم دربارهٔ بازه‌ای که مقدار واقعی θ با احتمال $۱ - \alpha$ به آن تعلق دارد حکمی احتمالاتی تهیه کنیم. یک برآورد بازه‌ای برای پارامتر θی جامعه، بازه‌ای به‌صورت $a < \theta < b$ است که در آن a و b به توزیع برآوردکنندهٔ $\hat{\theta}$ و به مقدار آن در ازای نمونهٔ موجود و به احتمالی که به این بازه تخصیص می‌یابد بستگی دارند. چون نمونه‌های مختلف ممکن به‌حجم n، مقادیر مختلف برای برآورد θ به‌وجود می‌آورند لذا به‌ازای این نمونه‌ها مقادیر متفاوتی برای a و b به‌دست می‌آیند. از آنجا که نمونه‌ها تصادفی‌اند این تصادفی بودن به مقادیر a و b نیز منتقل می‌شود، در نتیجه کل بازهٔ اطمینان، بازه‌ای تصادفی است. لذا بر پایهٔ توزیع $\hat{\theta}$ می‌توانیم

حکم کنیم که با احتمالی مفروض آیا این بازه، واقعاً پارامتر مجهول θی جامعه را در برمی‌گیرد یا نه. به‌عبارت دیگر، می‌توانیم از توزیع $\hat{\theta}$ استفاده کنیم و a و b را طوری بیابیم که با احتمالی برابر $1 - \alpha$، مقدار θ بین a و b بیفتد. بازهٔ $(a,\ b)$ را بازهٔ اطمینان $100(1-\alpha)$ درصد و $1-\alpha$ را ضریب اطمینان بازه و a و b را حدود اطمینان می‌نامند. برای یادآوری مطلب با طرح نمونه‌گیری که در آمار ریاضی دیده‌اید به ذکر مثالی می‌پردازیم.

مثال ۶.۱ فرض کنید جامعه‌ای بزرگ با میانگین مجهول $\overline{Y}_N$ داریم که انحراف معیار آن جامعه، σ، معلوم است. نمونه‌ای تصادفی (به‌گونه‌ای که در آمار ریاضی دیده‌اید) به‌حجم n از این جامعه می‌گیریم. می‌خواهیم برآوردی بازه‌ای برای $\overline{Y}_N$ مجهول با ضریب اطمینان ۹۵ر۰ به‌دست آوریم.

میانگین این نمونه را $\overline{Y}_n$ می‌گیریم، $\overline{Y}_n$ یک متغیر تصادفی است. تعداد $\overline{Y}_n$های ممکن با احتساب تکرارها، برابر با $\binom{N}{n}$ است. چون N بزرگ فرض شده است، با توجه به زیاد بودن بسیار تعداد $\overline{Y}_n$ها و معلوم بودن انحراف معیار آن که تقریباً برابر با $\frac{\sigma}{\sqrt{n}}$ است، مطابق قضیهٔ حدی مرکزی $\overline{Y}_n$ با میانگین $\overline{Y}_N$ و انحراف معیار $\frac{\sigma}{\sqrt{n}}$ دارای توزیع تقریبی نرمال است، و یا به عبارت دیگر

$$z = \frac{\overline{Y}_n - \overline{Y}_N}{\frac{\sigma}{\sqrt{n}}} \sim N(0,\ 1)$$

با مراجعه به جدول توزیع نرمال استاندارد می‌بینیم که $\Pr(Z \leq 1{,}96) \simeq 0{,}975$. پس می‌توانیم بنویسیم

$$\Pr(-1{,}96 < Z < 1{,}96) = 0{,}95$$

یا

$$\Pr\left(-1{,}96 < \frac{\overline{Y}_n - \overline{Y}_N}{\sigma/\sqrt{n}} < 1/96\right) = 0{,}95$$

یا با احتمال ۹۵ر۰، بازهٔ

$$\left(\underbrace{\overline{Y}_n - 1{,}96\frac{\sigma}{\sqrt{n}}}_{a} \quad , \quad \underbrace{\overline{Y}_n + 1{,}96\frac{\sigma}{\sqrt{n}}}_{b}\right)$$

شامل مقدار $\overline{Y}_N$ مجهول جامعه خواهد بود. چون برای هر نمونه، $\overline{Y}_n$ بخصوصی به‌دست می‌آید لذا در حالت کلی، برای هر نمونه یک بازهٔ $(a,\ b)$ی متفاوت از بازه‌های نمونه‌های دیگر به‌دست می‌آید. انتظار ما این است که اگر ۱۰۰ نمونه به‌حجم n از جامعهٔ مورد نظر بگیریم و برای هر کدام، بازهٔ اطمینان بالا را حساب کنیم ٪۹۵ این بازه‌ها مقدار واقعی $\overline{Y}_N$ را شامل باشند. ۵

در روشهای مختلف نمونه‌گیری معمولاً دربارهٔ توزیع جامعه اطلاعی در دست نیست، لذا از توزیع $\overline{Y}_n$ بی‌اطلاعیم، پس براساس این توزیع مجهول نمی‌توانیم بازهٔ اطمینان برای میانگین جامعه بیابیم. اما وقتی حجم جامعه بزرگ است می‌توان برحسب مورد توزیع تقریبی $\overline{Y}_n$ را نرمال یا استودنت گرفت و بازهٔ اطمینانی برای میانگین جامعه یافت.

۸. میانگین توان دوم خطا. وقتی پارامتر θی جامعه مجهول است، همیشه نمی‌توان به‌کمک نمونه، برآوردکننده‌ای نااریب برای آن یافت، و ممکن است $\hat{\theta}$ی حاصل از نمونه برآوردکننده‌ای اریب باشد. لذا $E(\hat{\theta})$ برابر با θ نبوده و برابر با مقدار مثلاً m است. اگر اریبی را با B نشان دهیم

$$B = E(\hat{\theta}) - \theta = m - \theta$$

برای مقایسهٔ دو برآوردکنندهٔ متفاوت θ، با اریبیهای مختلف، معمولاً از ملاک میانگین توان دوم خطا (MSE)* استفاده می‌کنند. منظور از خطا مقدار $\hat{\theta} - \theta$ است که مفهوم آن با اریبی یکی نیست، امید توان دوم این مقدار را $\text{MSE}(\hat{\theta})$ می‌نامیم، پس می‌توانیم بنویسیم

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^۲ = E[(\hat{\theta} - m) + (m - \theta)]^۲ \\ &= E(\hat{\theta} - m)^۲ + (m - \theta)^۲ + ۲(m - \theta)E(\hat{\theta} - m)\end{aligned}$$

توجه دارید که در عبارتهای بالا، تنها $\hat{\theta}$ متغیر تصادفی است. چون میانگین $\hat{\theta}$ را m فرض کردیم بنابراین در آخرین عبارت سمت راست، مقدار $E(\hat{\theta} - m)^۲$ برابر با واریانس $\hat{\theta}$ است. در جملهٔ سوم، $E(\hat{\theta} - m)$ میانگین انحرافهای $\hat{\theta}$ از میانگین آن است که برابر با صفر است، پس

$$\text{MSE}(\hat{\theta}) = \text{V}(\hat{\theta}) + B^۲ = \text{توان دوم اریبی + واریانس برآورد} \qquad (۲۹.۱)$$

معمولاً MSE را به‌عنوان ملاک دقت برآوردکننده به‌کار می‌برند. هرچه واریانس برآوردکننده کمتر و اریبی آن نیز کوچکتر باشد میانگین توان دوم خطاهایی که رخ می‌دهند کمتر است. از دو برآوردکننده، آنکه MSEی کمتری دارد بهتر است. دو برآوردکننده‌ای که MSE برابر دارند هم ارزند. اما این کاملاً دقیق نیست، زیرا توزیعهای فراوانی $\hat{\theta} - \theta$، به‌ازای حجمهای مختلف نمونه برای دو برآوردکننده، اگر اریبیهای متفاوت داشته باشند، یکسان نیستند. از آخرین رابطه نتیجه می‌شود که برای برآوردکننده‌های نااریب، MSE برابر با واریانس برآوردکننده است.

۹. اعتبار تقریب نرمال. در نظریهٔ احتمال دربارهٔ توزیع میانگین نمونهٔ تصادفی پژوهشهای زیادی شده است. ثابت می‌کنند که در هر جامعه با انحراف معیار متناهی، توزیع متغیر تصادفی

* MSE نمادی برای Mean Square Error، به‌معنای میانگین توان دوم خطاست.

میانگین نمونه، وقتی تعداد نمونه‌های ممکن زیاد شود به توزیع نرمال می‌گراید. برای نمونه‌گیری بدون جایگذاری که بعداً از آن صحبت خواهیم کرد، هایک[1] (۱۹۶۰) شرایط لازم و کافی برای گرایش توزیع میانگین نمونه به توزیع نرمال را ارائه کرده است. اردیش[2]، رن‌یی[3]، و مادو[4] کارهای او را دنبال کرده‌اند. این شرطها در اکثر موارد برقرارند. برای مطالعهٔ این شرایط علاقه‌مندان را به کتاب تکنیکهای نمونه‌گیری تألیف کوکران[5] ارجاع می‌دهیم.

هروقت فرض نرمال بودن توزیع $\overline{Y}_n$ پذیرفته شود محاسبهٔ بازهٔ اطمینان برای $\overline{Y}_N$ جامعه میسر است. در اینجا سؤالی که مطرح است این است که حجم نمونه باید تا چه حد بزرگ باشد، تا فرض نرمال بودن توزیع میانگین، فرضی معقول به‌حساب آید. درواقع پاسخ این است که قاعدهٔ کلی مطمئنی وجود ندارد. برای جامعه‌هایی که می‌توان از روی بافتنمای نمونه دریافت که چاولگی مثبت بارزی دارند، قاعده‌ای که در اکثر مواقع مفید است انتخاب n از روی نابرابری

$$n > ۲۵G_۱^۲ \qquad (۳۰.۱)$$

است، که در آن $G_۱$، معیار چاولگی فیشر است و به‌صورت زیر بیان می‌شود

$$G_۱ = \frac{۱}{\sigma^۳}E(Y_i - \overline{Y}_N)^۳ = \frac{۱}{N\sigma^۳}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۳$$

که در آن، N حجم جامعه است. با انتخاب n از روی نابرابری بالا، تحقیق شده است که یک حکم احتمالاتی بازهٔ اطمینان ۹۵ درصد، کمتر از ۶٪ موارد غلط درمی‌آید. چون $G_۱$ پارامتری مربوط به جامعه بوده و مجهول است، در عمل آن را از روی نمونه‌ای مقدماتی برآورد می‌کنند و برای تعیین برآورد کران پایین n به‌کار می‌برند.

۱۰. ضریب تغییرات. مفهوم دیگری که غالباً در بحث روشهای نمونه‌گیری به‌خصوص در برآورد نسبتی و رگرسیونی مطرح می‌شود مفهوم ضریب تغییرات است. در یک جامعه، معیار پراکندگی، میزان انحرافهای واحدهای جامعه از میانگینشان است که از روی واریانس محاسبه می‌شود. در بسیاری از موارد برای اهداف مقایسه‌ای، انحرافها را برحسب مقدار نسبی آنها در نظر می‌گیرند. ضرورت این مطلب در هنگام مقایسهٔ پراکندگی دو جامعه آشکار می‌شود. مثلاً اگر در اندازه‌گیریهای مکرر طول یک شیء، میانگین طولها ۱۰ متر و انحراف معیار اندازه‌ها ۱ر۰ متر باشد و اگر در اندازه‌گیریهای مکرر قطر یک صفحهٔ بسیار نازک فلزی میانگین اندازه‌ها ۳ر۰ میلیمتر و انحراف معیار اندازه‌ها ۱۵ر۰ میلیمتر باشد نمی‌توان با مقایسهٔ دو انحراف معیار دربارهٔ اینکه پراکندگی کدام یک از دو دنبالهٔ اندازه‌ها بیشتر است قضاوت کرد، زیرا ممکن است به‌دلیل بزرگ بودن طول شیء، یک انحراف از میانگین به ظاهر بزرگ در اندازه‌گیریهای طول شیء، در مقایسهٔ با انحراف از میانگین به ظاهر کوچکی در اندازه‌گیریهای قطر صفحه، خیلی کوچک باشد، همین مطلب است که محاسبهٔ مقدار نسبی

1. Hájek 2. Erdös 3. Rényi 4. Madow 5. Cochran

انحرافها را ضروری می‌سازد. لذا اگر در محاسبهٔ واریانس به‌جای منظور کردن $|Y_i - \overline{Y}_N|$ها از $|\frac{Y_i - \overline{Y}_N}{\overline{Y}_N}|$ استفاده کنیم و این واریانس را با $C^۲$ نشان دهیم، فرمول واریانس به‌صورت زیر درمی‌آید

$$C^۲ = \frac{۱}{N}\sum_{i=۱}^{N}\left(\frac{Y_i - \overline{Y}_N}{\overline{Y}_N}\right)^۲$$
$$= \frac{۱}{N\overline{Y}_N^۲}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲$$

اما $\sigma^۲ = \frac{۱}{N}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲$، پس

$$C^۲ = \frac{\sigma^۲}{\overline{Y}_N^۲}$$

این ملاک پراکندگی، اثر انحرافهای نسبی را در پراکندگی منعکس می‌کند. $C^۲$ را واریانس نسبی جامعهٔ یا واریانس نسبی توزیع متغیر تصادفی می‌نامند و

$$C = \frac{\sigma}{\overline{Y}_N} \qquad (۳۱.۱)$$

را ضریب تغییرات جامعه می‌گویند. این ضریب امکان مقایسه تغییرات دو جامعه را میسر می‌سازد. در مثال اندازه‌گیریهای مکرر طول شیء و اندازه‌گیریهای مکرر قطر صفحهٔ فلزی، دو ضریب تغییرات به‌ترتیب $C_۱ = \frac{۰٫۱}{۱۰} = ۰٫۰۱$ و $C_۲ = \frac{۰٫۰۱۵}{۰٫۳} = ۰٫۰۵$. می‌بینیم که تغییرپذیری اندازه‌گیریهای قطر صفحهٔ فلزی بیشتر است، یعنی به‌طور نسبی عمل اندازه‌گیری قطر صفحهٔ فلزی نادقیقتر بوده است.

ضریب تغییرات را غالباً برحسب درصد بیان می‌کنند. به مثال زیر توجه کنید.

مثال ۷.۱ میانگین وزن گروهی از دانشجویان $\overline{Y}_N = ۷۰$ کیلوگرم است. انحراف معیار وزن این دانشجویان $\sigma = ۵٫۲۵$ کیلوگرم است. ضریب تغییرات وزنها

$$C = \frac{۵٫۲۵}{۷۰} = ۰٫۰۷۵$$

در این صورت می‌گوییم ضریب تغییرات ۷٫۵ درصد است. ▲

ویژگی مهم C این است که مقدار آن در جامعه‌هایی که تغییراتی طبیعی دارند تقریباً ثابت است. مثلاً ضریب تغییر وزن دانشجویان یک کلاس در طول سال خیلی کم تغییر می‌کند و یا تغییر Cی میزان محصول گندم یک مزرعه بزرگ از جریبی به جریب دیگر خیلی کم است و یا تغییر Cی داده‌های یک آزمایش، وقتی چندین بار تکرار می‌شود فوق‌العاده اندک است. اگر جامعه

پیوسته باشد و یا اگر برای N بزرگ واحدهای جامعه Y_1، Y_2، . . . ، Y_N باشند، همان‌طور که دیدیم وقتی Y، متغیر تصادفیی باشد. که این مقادیر را می‌پذیرد، آن‌گاه

$$C(Y) = \frac{\sigma_Y}{\overline{Y}_N}$$

در این صورت اگر نمونه‌ای به‌حجم n به تصادف از جامعه بگیریم و $\overline{Y}_n$ میانگین نمونه باشد، $\overline{Y}_n$ متغیری تصادفی است و داریم

$$E(\overline{Y}_n) = \overline{Y}_N \quad , \quad \mathrm{V}(\overline{Y}_n) = \frac{\sigma_Y}{\sqrt{n}}$$

لذا برای دنبالهٔ $\overline{Y}_n$های ممکن

$$C(\overline{Y}_n) = \frac{\sqrt{\mathrm{V}(\overline{Y}_n)}}{E(\overline{Y}_n)} = \frac{\sigma_Y}{\overline{Y}_N\sqrt{n}} = C(Y)\cdot\frac{۱}{\sqrt{n}} \qquad (۳۲.۱)$$

یعنی در چنین جامعه‌هایی، ضریب تغییرات میانگین نمونه‌ای به‌حجم n برابر با $\frac{۱}{\sqrt{n}}$ ضریب تغییرات جامعه است.

مثال ۸.۱ اگر بخواهیم انحراف معیار $\overline{Y}_n$ نمونه‌ای تصادفی به‌حجم n برابر با ۱٫۵ درصد $\overline{Y}_n$ باشد حجم نمونه را چقدر باید انتخاب کنیم؟ حجم جامعه را بزرگ و نمونه را تصادفی گرفته‌ایم.

وقتی می‌گوییم در یک جامعه مثلاً $C = $ ۰٫۳، یعنی انحراف معیار جامعه برابر ۰٫۳ میانگین این جامعه است. در این مثال $C(\overline{Y}_n) = $ ۰٫۰۱۵، پس

$$C(\overline{Y}_n) = ۰٫۰۱۵ = \frac{C(Y)}{\sqrt{n}} = \frac{۰٫۳}{\sqrt{n}}$$

لذا

$$\sqrt{n} = \frac{۰٫۳}{۰٫۰۱۵} = ۲۰ \quad , \quad n = ۴۰۰$$

Δ

ازکاربردهای عمدهٔ ضریب تغییرات، مقایسهٔ تغییرات دو جامعه و یا مقایسهٔ تغییرات دو مشخصه از واحدهای یک جامعه است. اگر در جامعهٔ $(X_i,\ Y_i)$، $i = ۱,\ ۲, \ldots,\ N$، X_i معرف قد فرد iام و Y_i معرف وزن این فرد باشد، بدیهی است که σ_X معیاری برای میزان تغییر قد در این جامعه است که برحسب واحد اندازهٔ قد که سانتیمتر است بیان می‌شود. معیار تغییر وزن در این جامعه σ_Y، و برحسب کیلوگرم است. اگر بخواهیم بدانیم که تغییرات کدام یک از دو مشخصه بیشتر

است مقایسهٔ دو مقدار σ_X و σ_Y با واحدهای مختلف، منطقاً معرّف چگونگی اختلاف تغییرات دو مشخصه نیست. در حالی که اگر ضریب تغییرات دو صفت، یعنی $C_X = \frac{\sigma_X}{\overline{X}_N}$ و $C_Y = \frac{\sigma_Y}{\overline{Y}_N}$ را که بدون واحد بوده و در عین حال تغییرات نسبی را نشان می‌دهند در نظر بگیریم، مقایسهٔ آنها نشان می‌دهد که تغییرات کدام از یک از دو صفت بیشتر است. معمولاً با نمونه‌گیری برآوردهای $\hat{C}_X$ و $\hat{C}_Y$ را به‌دست می‌آورند و آنها را با هم مقایسه می‌کنند.

مثال ۹.۱ میانگین نمرات درس آمار دانشجویان یک کلاس ۱۶ و واریانس این نمرات ۴ است. میانگین وزن همین دانشجویان برابر با ۶۸ کیلوگرم و واریانس وزنها برابر با ۹ است. تغییرات کدام یک از این دو مشخصه بیشتر است؟

مطابق آنچه گفتیم اگر نمرهٔ درس آمار را با X و وزن دانشجویان را با Y نشان دهیم

$$\overline{X}_N = ۱۶ \ , \ \sigma_X = ۲ \ , \ \overline{Y}_N = ۶۸ \ , \ \sigma_Y = ۳$$

بنابراین

$$C_X = \frac{۲}{۱۶} = \frac{۱}{۸}, \quad C_Y = \frac{۰۳}{۶۸}$$

چون $C_X > C_Y$، تغییرات نمرات درس آمار بیش از تغییرات وزنهاست گرچه به‌ظاهر σ_X از σ_Y کمتر است. ▲

۱۰.۱ روشهای گردآوری داده‌ها

متداولترین روشهای گردآوری داده‌ها در بررسیهای نمونه‌ای عبارت‌اند از:
مصاحبه‌های حضوری، و مصاحبه‌های تلفنی. تجربه نشان داده است که اگر در این روشها مصاحبه‌کننده آموزش لازم را دیده باشد و مکالمهٔ تلفنی یا حضوری از پیش طرحریزی شود و در تمام فرایند، از نظر روانشناسی ملاحظات لازم صورت گیرد نرخ پاسخدهی از ۶۰ تا ۷۵ درصد است. این نرخ می‌تواند حتی بالاتر از این میزان نیز باشد. پرسشنامه‌ای پستی که برای افراد منتخب با روش نمونه‌گیری فرستاده می‌شود می‌تواند نتایجی رضایتبخش به‌دست دهد، اما در حالت کلی نرخ پاسخ برای این نوع گردآوری داده آن‌قدر پایین است که نمی‌توان به گزارش استنباط حاصل اعتماد داشت. غالباً اطلاعات لازم را به‌جای مصاحبهٔ تلفنی یا ارسال پرسشنامه باید با مشاهدهٔ مستقیم فراهم کرد. در زیر چهار نوع گردآوری داده را مورد بحث قرار می‌دهیم.

۱. **مصاحبه‌های فردی.** داده‌ها غالباً از طریق مصاحبه‌های فردی به‌دست می‌آیند. مثلاً می‌توان از مصاحبه با افراد منتخب با روش نمونه‌گیری، نمونه‌ای را از نظر جامعه دربارهٔ انتشار اوراق قرضه به‌دست آورد. شیوهٔ کار، مستلزم تعیین مصاحبه‌گرانی برای مطرح کردن سؤالها و ثبت پاسخهاست.

مزیت عمدهٔ این روش آن است که معمولاً افراد در رویارویی با مصاحبه‌گر به سؤالها پاسخ می‌دهند. به‌علاوه مصاحبه‌گر می‌تواند واکنشهای خاص پاسخ‌دهنده را ملاحظه و سوءتفاهماتی که ممکن است در مورد سؤال پیش آید مرتفع کند. محدودیتهای عمده مصاحبهٔ فردی (علاوه بر هزینه) در ارتباط با مصاحبه‌گرهاست. اگر مصاحبه‌گر آموزش کافی ندیده باشد ممکن است با طرح نادرست برخیٔ سؤالها، در داده‌های نمونه‌ای، اریبی ایجاد کند. هر حرکت، هر حالت چهره، و یا نحوهٔ بیان مصاحبه‌گر بر پاسخ سؤال اثر می‌گذارد. مثلاً سؤالی نظیر «آیا شما هم موافق انتشار اوراق قرضه هستتید؟» ممکن است به جواب مثبت مصاحبه‌شونده منجر شود. بدیهی است خطاهای ممکن در ثبت پاسخها هم نتایج استنباط از پاسخها را دچار خطا می‌کند. برای رفع این خطاها توصیه‌های لازم در ادبیات آماری وجود دارند.

۲. مصاحبه‌های تلفنی. می‌توان اطلاعات مربوط به نمونه را از طریق مصاحبه‌های تلفنی تهیه کرد. مصاحبه‌گر می‌تواند از دفتر تلفن شهری به تصادف هر شماره‌ای را انتخاب کند. بدیهی است داده‌های حاصل از طریق تلفن با هزینه‌ای کم همراه است، زیرا هزینه‌های رفت و آمد مصاحبهٔ فردی را ندارد. مشکل عمدهٔ بررسیهای تلفنی، استقرار چارچوبی است که متناظر با جامعه باشد. دفتر تلفن شامل نمره‌های بسیاری است که به خانواده‌ها تعلق ندارند و بسیاری از خانواده‌ها هستند که یا تلفن ندارند و یا شمارهٔ تلفن آنها در دفتر تلفن ثبت نشده است. یک راه حل نسبی مشکل این است که بدون توجه به دفتر تلفن، شماره‌گیریها با ارقام تصادفی انجام شوند. این شماره‌گیری، شماره‌های تلفنی را هم که در دفتر تلفن نیستند در چارچوب منظور می‌کند.

مصاحبه‌های تلفنی باید معمولاً کوتاه باشند، زیرا افرادی که مصاحبه می‌شوند در تماسهای تلفنی به‌سرعت بی‌حوصله شده و این بی‌حوصلگی به جوابهای اریب منجر خواهند شد. با سؤالهایی که دقیقاً طراحی شده‌اند و با مصاحبه‌گرهای آموزش دیده، مصاحبه‌های تلفنی غالباً موفقیت‌آمیزند.

۳. پرسشنامه. روش دیگر مفید گردآوری داده، پرسشنامه است که به‌وسیلهٔ پاسخ‌دهنده تکمیل می‌شود. این پرسشنامه به‌طریق پستی برای افرادی که به‌عنوان واحدهای نمونه انتخاب شده‌اند ارسال می‌شود. پرسشنامه باید با دقتی خاص طرحریزی شود که رغبت افراد را به دادن پاسخ برانگیزد.

پرسشنامه‌ها نیازی به مصاحبه‌گر ندارند و لذا با صرفه‌جویی در هزینه همراه‌اند. در عوض، این صرفه‌جویی در هزینه، موجب نرخ پایین پاسخهاست، زیرا بسیاری از پرسشنامه‌ها بی‌پاسخ می‌مانند. بی‌پاسخی اصولاً یک مشکل عمده هر روش گردآوری داده است، زیرا پاسخهای کم، درواقع حجم نمونه را کاهش می‌دهند و در نتیجه در حاصل کار اریبی به‌وجود می‌آورند. برای کم کردن تعداد پرسشنامه‌های بی‌پاسخ معمولاً به ارسال مجدد پرسشنامه می‌پردازند و در بار دوم تعدادی پاسخ دریافت می‌شود که از اریبی می‌کاهند. گاهی برای اخذ پاسخ در بار دوم از طریق تلفن و یا مصاحبهٔ حضوری اقدام می‌کنند.

۴. مشاهدهٔ مستقیم. چهارمین روش گردآوری داده، روش مشاهدهٔ مستقیم است. مثلاً اگر بخواهیم تعداد ماشینهایی را که از ساعت ۴ تا ۶ عصر از جاده‌ای خاص عبور می‌کنند برآورد کنیم، باید فردی را مأمور کنیم که ماشینها را در نقطه‌ای از این جاده در ساعتهای مورد نظر بشمارد. در صورت امکان می‌توان از دستگاههای شمارشگر الکترونیکی نیز استفاده کرد. بدیهی است هر مأمور ممکن است در شمارش ماشینها دچار خطا شود.

از مشاهدهٔ مستقیم در اکثر بررسیهایی که مستلزم اندازه‌گیری روی واحدهای نمونه نیست استفاده می‌کنند. مثلاً زیست‌شناسان دنیای وحش، برای برآورد حجم جامعه‌های پرندگان به شمارش آشیانه‌ها و تخمها می‌پردازند.

یکی از مزایای مشاهدهٔ مستقیم آن است که نتایج، تحت تأثیر رفتار یا واکنش پاسخگویان نیستند. اطلاعات بهداشتی را می‌توان از سوابق بیمارستان به‌دست آورد. اطلاعات مربوط به درآمدها را می‌توان از اداره‌های مالیاتی کسب کرد. این روش گرچه ممکن است وقتگیر باشد ولی نتایج حاصل غالباً قابل اطمینان‌اند.

۱۱.۱ انواع عدم حتمیتها در داده‌ها

همهٔ داده‌ها، چه حاصل از سرشماری و چه حاصل از نمونه‌گیری، با انواع مختلفی از عدم حتمیت همراه‌اند. می‌توان میزان عدم حتمیتها در داده‌ها را، پس از تشخیص وجود آنها، با اصلاح گام‌به‌گام در بررسیهای آتی، کاهش داد. درواقع طرح نمونه‌گیری، تلاش برای رسیدن به یک تعادل بهینه بین انواع مختلف عدم حتمیتهاست. معمولاً سه نوع عدم حتمیت در داده‌ها وجود دارند:

۱. عدم حتمیت نوع ۱ حاصل نقصها و محدودیتهای ساختاری چارچوب، پرسشنامه و روش طراحی پرسشهاست. پاسخ به یک پرسشنامه و یا ثبت هر اندازه‌ای که با وسیله‌ای اندازه‌گیری می‌شود، تنها پاسخی به نوع سؤال مطروحه است. بنابراین نقصهای ممکن در پرسشنامه و یا نقصهای نحوهٔ بیان شفاهی سؤالها و استفاده از روشهای نامناسب، ممکن است موجب تفهیم ناکامل هدف سؤال‌شده و موجب به‌دست آمدن داده‌های ناقص شوند. این نوع عدم حتمیت مستقل از حجم یا نوع نمونه است و به‌وسیلهٔ موشکافی در داده‌های حاصل کشف نمی‌شوند و یا با محاسبهٔ خطای معیار یا سایر محاسبه‌های آماری آشکار نخواهند شد. بعضی از خطاهای نوع لول به صورتهای زیر رخ می‌دهند:

الف) ممکن است در چارچوب، برخی از بخشهای جامعه نادیده گرفته شوند.

ب) پرسشنامه یا روش بیان مفاد آن ممکن است در استخراج اطلاعاتی که بعداً نیاز به آنها مشخص می‌شود توانا نباشند. پرسشنامه ممکن است شامل تعریفها و سؤالهای نارسا باشند. تفاوت تأثیر مصاحبه‌کننده‌ها که حاصل ویژگیهای روانی، جنس، نژاد، و سن است ممکن است بر میزان درستی پاسخها اثر گذارد.

ج) استفاده از تلفن یا پست ممکن است پاسخهایی غیر از پاسخهای حاصل از مصاحبهٔ حضوری را به‌دست دهند.

د) داوریهای کدگذارها ممکن است با هم متفاوت باشند.

ه) وقت و تاریخ بررسی می‌تواند بر میزان درستی داده‌های حاصل از پاسخها اثر بگذارد.

۲. عدم حتمیتهای نوع ۲ شامل خطاها و نقصهای عملیاتی و محاسباتی است، نظیر

الف) اریبی حاصل از پرسشنامه‌های بی‌پاسخ.

ب) اطلاعاتی که از اِعمال روشی خاص برای خنثی‌کردن پرسشنامه‌های بی‌پاسخ یا پرسشنامه‌هایی که ناخوانا هستند حاصل می‌شوند.

ج) وجود خطای عمده در محاسبات.

۳. عدم حتمیت نوع ۳ به‌وسیلهٔ تغییرات تصادفی حاصل می‌شوند.

نمونه‌های تصادفی مکرر که از یک چارچوب به‌دست می‌آیند نتایج مختلفی به‌دست می‌دهند. این عدم حتمیتها در هر روش و طرحی که برای نمونه‌گیری در نظر گرفته می‌شود با قانونهای احتمال بررسی می‌شوند. علاوه بر این تغییرات تصادفی، تغییرات تصادفی ذاتاً ناهمبسته و ناسازگاری وجود دارند که حاصل طبیعت تغییرپذیری تصادفی بررسی‌کنندگان است. خطای معیاری که از داده‌ها نتیجه می‌شود شامل آمیزه‌ای از اثرهای آمیختهٔ همه نوع تغییر تصادفی، از جمله تغییرات تصادفی خود بررسی‌کننده و بین بررسی‌کنندگان، ناظران، کدگذاران و غیره است، که با طرح مناسب نمونه‌گیری امکان جدا ساختن بعضی از تغییرات تصادفی وجود دارد.

به‌طور کلی خطاهایی که در بالا به آنها اشاره شد به دو دسته تقسیم می‌شوند. چون نمونه نمی‌تواند اطلاع کامل دربارهٔ جامعه بدهد لذا خطایی در برآورد پارامترهای جامعه رخ خواهد داد که این نوع خطا را خطای نمونه‌گیری می‌نامند. خطای نمونه‌گیری را می‌توان با طرح مناسب نمونه‌گیری که موضوع این کتاب است تا حدی کنترل کرد. انواع دیگر خطاها را که کنترل آن مشکل است خطای غیرنمونه‌گیری می‌نامند.

نکتهٔ مهمی که باید متذکر شویم آن است که در همهٔ فصول آینده، در روشهای مختلف نمونه‌گیری، فرمولهایی که ارائه می‌شوند مبتنی بر این فرض‌اند که وقتی واحدهای نمونه مشخص می‌شوند و به اندازه‌گیری آنها می‌پردازیم و براساس این اندازه‌ها فرمولها را به‌صورت ریاضی و احتمالاتی بیان می‌کنیم خطای اندازه‌گیری نداریم و تنها خطای موجود خطای نمونه‌گیری است.

۱۲.۱ طرّاحی پرسشنامه

هدف عمدهٔ طرح نمونه‌گیری، مینیمم کردن خطاهای غیرنمونه‌گیریی است که ممکن است رخ دهند. اگر بررسی، شامل کسب اطلاعی از افراد باشد، باید از وقوع خطاهای غیرنمونه‌گیری زیادی جلوگیری شود. این جلوگیری غالباً با طرح یک پرسشنامه حساب‌شده امکان‌پذیر است. ملاحظات عمده‌ای که در ساختن یک پرسشنامه باید مورد نظر باشند به‌شرح زیرند:

الف) ترتیب سؤالها. ترتیب طرح سؤالها در پرسشنامه از اهمیتی خاص برخوردار است، زیرا ترتیب سؤالها بر پاسخها اثر می‌گذارد. با مثالی می‌توان این نکته را روشن کرد.

در پرسشنامه‌ای دو سؤال زیر مطرح شده‌اند:

I. آیا فکر می‌کنید که آمریکا باید به خبرنگاران کمونیست سایر کشورها اجازه دهد که به آمریکا بیایند و مشاهدات خود را برای روزنامه‌های کشورشان بفرستند؟

II. آیا فکر می‌کنید که کشورهای کمونیستی نظیر چین اجازه می‌دهند که خبرنگاران آمریکایی به چین بروند و مشاهدات خود را برای روزنامه‌های آمریکایی بفرستند؟

این سؤالها به‌ترتیب (I, II) در ۱۹۸۰ به‌بررسی گذاشته شده‌اند و ۵۴٫۷٪ پاسخ‌دهندگان به I پاسخ آری و ۶۳٫۷٪ آنها به II پاسخ آری داده‌اند. بار دیگر همین دو سؤال به‌ترتیب (II, I) به بررسی گذاشته شده‌اند و ۷۴٫۶٪ پاسخ‌دهندگان به I پاسخ آری و ۸۱٫۹٪ پاسخ‌دهندگان به II پاسخ آری داده‌اند. بدین‌ترتیب شواهد حاکی از این است که اگر سؤال II اول مطرح شود پاسخ‌دهندگان را در وضعیتی قرار می‌دهد که پاسخ آری بیشتری به سؤال I می‌دهند. به‌عبارت دیگر، آنهایی که جواب آری به II می‌دهند وقتی اول مورد سؤال قرار می‌گیرند سعی دارند که سازگاری خود را با پاسخ آری به سؤال I نشان دهند. به مثال دیگری توجه کنید. دو سؤال زیر را از افراد منتخب یک جامعه می‌پرسیم

I. آیا مایل هستید با پرداخت مالیات بیشتر به بسط و توسعهٔ مدارس کمک کنید؟

II. آیا مایل هستید به‌دولت مالیات بیشتری بپردازید؟

اگر ابتدا سؤال II مطرح شود و سپس سؤال I، افراد بیشتری به II جواب آری می‌دهند تا وقتی ابتدا سؤال I مطرح شود، زیرا اگر سؤال I را ابتدا مطرح کنیم افرادی که از افزایش مالیات به‌دلیل توسعهٔ مدارس حمایت می‌کنند و به I جواب مثبت می‌دهند ممکن است در مورد سؤال دوم یعنی II فکر کنند که افزایش مالیات لزوماً به بسط مدارس تخصیص داده نمی‌شود و در نتیجه به سؤال II پاسخ منفی می‌دهند. اگر سؤال II اول مطرح شود افرادی که موافق افزایش مالیات‌اند و در عمق ذهن خود بدون آگاهی از سؤال اول، توجه به بسط مدارس دارند به سؤال II پاسخ مثبت می‌دهند. ملاحظه می‌شود که چگونه در دو مثال بالا ترتیب مطرح کردن سؤال در نتیجهٔ کار مؤثر است. طراح پرسشنامه باید ترتیب سؤالها را به‌گونه‌ای در نظر بگیرد که میزان اثرهای متقابل را به حداقل برساند، که این خود مستلزم بحثی تکمیلی است.

ب) انواع سؤالها. معمولاً در پرسشنامه‌ها دو نوع سؤال، چند پاسخی و آزاد، مطرح می‌شوند. در سؤال چند پاسخی، برای هر سؤال دو یا چند پاسخ درج می‌شود و پاسخگو باید از بین آنها پاسخی را برحسب باوری که دارد انتخاب کند. پاسخهای مندرج باید روشن بوده و دو به دو نامتداخل و ناسازگار باشند و اگر سؤال، پاسخهای متعددی را ایجاب می‌کند همهٔ حالتهای ممکن اصلی را شامل باشند. عیب عمدهٔ این نوع سؤال آن است که پاسخهای مندرج، القاکننده هستند و اگر درج پاسخ به عهدهٔ پاسخگو بود ممکن بود به‌صورتی غیر از پاسخهای مندرج پاسخ دهد.

نوع دیگر، سؤال آزاد-پاسخ است. در این نوع سؤال، تهیهٔ هر نوع پاسخی به‌عهدهٔ پاسخگو است. سؤال زیر از این نوع است:

به‌عقیدهٔ شما، مزایای کار معلمی چه هستند؟

بدیهی است پاسخهایی که برای چنین سؤالی دریافت می‌شوند متعددند و لاجرم باید رده‌بندی شوند.

این رده‌بندی کار مشکلی است زیرا باید همهٔ پاسخها را با هم مقایسه کرد و پاسخهای تقریباً مشابه را مشخص کرد. تحلیل این نوع پاسخها نیز آسان نیست. بدین دلایل معمولاً سؤالهای آزاد-پاسخ در مطالعه‌های مربوط به جامعه‌های کوچک به‌کار می‌روند. گاهی نیز وقتی با بررسی قبلی، تعداد پاسخهای ممکن محدود است از این سؤالها استفاده می‌شود. گاهی قالب سؤال چندپاسخی و آزاد-پاسخ با استفاده از واژهٔ «هیچکدام» ادغام می‌شوند در نتیجه پاسخگو می‌تواند یا به یکی از پاسخهای مندرج اشاره کند، و یا پاسخ خود را به‌صورت «هیچکدام» مشخص کند. معمولاً وقتی پاسخ در ردهٔ «هیچکدام» است با سؤالهای بعدی اطلاعاتی دربارهٔ نظر او کسب می‌شود.

ج) روشنی سؤالها. سؤالها باید به‌گونه‌ای عبارت‌بندی شوند که هر سؤال برای هر پاسخگو مفهومی واحد را القا کند. نباید از اصطلاحات و واژه‌هایی استفاده کرد که تعریف آنها برای عموم مشخص نیست و جنبهٔ تکنیکی و علمی خاص دارند. جمله‌ها باید کوتاه باشند. سؤالات طولانی، همیشه با تعبیر نادرست پاسخگو، نتایج کار را مخدوش می‌سازد.

د) توجه به جنبه‌های روانی پاسخگو. وقتی سؤالها جنبهٔ کیفی دارند و این جنبه‌های کیفی در ارتباط با شخص پاسخگو یا خانواده اوست، بسیاری از پاسخگویان به‌صورتی اغراق‌آمیز پاسخ می‌دهند و برای ارضای غرور خود پاسخهایی اریب می‌دهند. در این موارد می‌توان با طرح سؤالهای خاص این اریبیها را به حداقل رساند. بدین معناکه برای حفظ اعتبار و شخصیت پاسخگو می‌توان برای رسیدن به پاسخ صحیح از سؤالهایی به روش غیرمستقیم استفاده کرد. مثلاً به‌جای اینکه بپرسیم «آیا تحصیلات دانشگاهی دارید؟» می‌توانیم سؤال کنیم که «در زمان ترک تحصیل چه کلاسی را به پایان رساندید؟»

ه) اجتناب از سؤالهای جهت‌دار. گاهی سؤالها به‌گونه‌ای هستند که به پاسخگو در دادن پاسخ جهت خاصی را القا می‌کند مثلاً این سؤال که «آیا شما هم با نظام جدید آموزش متوسطه موافق هستید؟» به دلیل وجود واژه‌های «هم» و «موافق» موافق بودن را در بسیاری از پاسخگویان القا می‌کند. در حالی که اگر سؤال به‌صورت «آیا شما هم با نظام جدید آموزش متوسطه مخالف هستید؟» مطرح شود در بسیاری از پاسخگویان مخالف بودن را القا می‌کند. این دو نوع سؤال، جهت دارند. باید از گنجاندن سؤالهایی بدین صورت در پرسشنامه خودداری کرد.

و) آزمون مقدماتی پرسشنامه. وقتی پرسشنامه‌ای تنظیم می‌شود باید آن را به مقیاس کوچک اجرا کرد تا به مشکلات پاسخگویان در تعبیر سؤالها پی‌برد، درست بودن ترتیب سؤالها و نامتداخل بودن آنها، انشای آنها و صراحت دستورالعملها را کنترل کرد و پس از رفع این نقایص احتمالی، پرسشنامه را در نمونه‌ای به‌حجم مورد نظر به‌کار برد.

تمرینها

۱ . جنگلبانی می‌خواهد در یک ناحیه از جنگل، تعداد درختهایی را که قطری بیش از ۳۰ سانتیمتر دارند برآورد کند. نقشه‌ای از ناحیه دردست است. دربارهٔ مسألهٔ انتخاب واحد نمونه‌گیری مناسب و چارچوب مناسب بحث کنید.

۲ . صنعتی خاص، کارگاههای متعددی در سراسر کشور دارد. مقام اجرایی این صنعت می‌خواهد دربارهٔ سیاست ادارهٔ کارگاهها از کارکنان نظرخواهی کند. دربارهٔ انتخاب واحدهای نمونه‌گیری چه پیشنهادی دارید؟ چه چارچوبی را پیشنهاد می‌کنید؟

۳ . فرض کنید می‌خواهیم میزان متوسط مصرف قند خانواده‌ها را در شهری برآورد کنیم. جامعه و واحد جامعه را تعریف کنید. اگر فهرستی از خانواده‌ها موجود نباشد ولی نقشه‌ای از شهر داشته باشیم که تمام بلوکها را مشخص کند در این صورت جامعه و واحد جامعه را تعریف کنید. اگر بخواهیم نمونه‌ای تهیه کنیم چارچوب نمونه‌گیری را پیشنهاد کنید.

۴ . فرض کنید متوسط وزن گروهی از دانشجویان $\overline{Y} = ۶۰$ کیلوگرم، انحراف معیار وزنها $\sigma. = ۳$ کیلوگرم باشد، ضریب تغییرات و واریانس نسبی را تعیین کنید.

۵ . وزن متوسط گروهی از دانشجویان $\overline{Y} = ۷۰$ کیلوگرم است. فرض کنید از داده‌های قبلی می‌دانیم که ضریب تغییرات معمولاً ۷٪ است. انحراف معیار وزن این گروه را برآورد کنید.

۶ . فرض کنید در جامعه‌ای ضریب تغییرات ۳۰٪ است. می‌خواهیم انحراف معیار $\overline{Y}_n$ برابر با ۳ درصد $\overline{Y}_N$ باشد در این صورت حجم نمونه را باید چقدر انتخاب کنیم؟ $\overline{Y}_n$، میانگین نمونه‌ای به‌حجم n و $\overline{Y}_N$، میانگین جامعه به‌حجم N است.

۷ . نمرات ۵ دانشجو در درسی به‌ترتیب ۶۰، ۶۵، ۷۰، ۷۰، ۹۵ است. اگر نمونه‌هایی به‌حجم ۲ انتخاب کنیم و $\overline{Y}_n$ متغیر تصادفی معرف میانگین نمونه‌ها باشد، ضریب تغییرات $\overline{Y}_n$ را به‌دست آورید.

۸ . وزنهای ۱۰ دانشجو به‌ترتیب برابرند با

$$۶۵,\ ۶۷,\ ۷۵,\ ۸۰,\ ۶۲,\ ۵۵,\ ۷۲,\ ۷۳,\ ۶۰,\ ۵۸$$

کیلوگرم و فشار خون آنها به‌ترتیب برابرند با

$$۱۲,\ ۱۳,\ ۱۳,\ ۱۴,\ ۱۲٫۵,\ ۱۲,\ ۱۳,\ ۱۳٫۵,\ ۱۳,\ ۱۴$$

تعیین کنید میزان تغییرات وزن بیشتر است یا فشار خون؟

۹ . در جامعه‌ای گشتاور مرکزی مرتبهٔ سوم برابر با ۱۵ برآورد شده است. برآورد انحراف معیار جامعه برابر ۲ است. در این جامعه که چاولگی زیاد دارد برای اینکه $\overline{Y}_n$، میانگین نمونه، توزیعی تقریباً نرمال داشته باشد حداقل حجم نمونه را چقدر باید انتخاب کرد؟

۱۰ . دو متغیر تصادفی X و Y مستقل از یکدیگرند. توزیع X، نرمال با میانگین ۱ و واریانس ۴ است. متغیر Y دارای توزیع نرمال استاندارد است. اگر $Z = XY$ باشد واریانس Z را بیابید.

۲

نمونه‌گیری تصادفی ساده

۰.۲ مقدمه

ساده‌ترین روش نمونه‌گیری احتمالاتی روش نمونه‌گیری تصادفی ساده است. گرچه این روش در کاربردهای فعلی زیاد مورد استفاده نیست ولی در آغاز بحث روشهای مختلف نمونه‌گیری، ذکر این روش ضروری است، زیرا، برای بیان و طرّاحی روشهای دقیقتر نمونه‌گیری، به‌درک این روش نیاز داریم. به‌همین دلیل گاهی این روش نمونه‌گیری را روش نمونه‌گیری مادر می‌نامند. غالباً برای سهولت، این روش را فقط روش نمونه‌گیری تصادفی می‌خوانند. در این فصل روش نمونه‌گیری تصادفی بدون جایگذاری و با جایگذاری را شرح می‌دهیم و از این روشها برای برآورد میانگین جامعه، واریانس جامعه و همچنین برآورد نسبت دارندگان مشخصه‌ای معلوم در جامعه استفاده می‌کنیم.

۱.۲ تعریف نمونه‌گیری تصادفی ساده بدون جایگذاری

روش نمونه‌گیری تصادفی ساده، روش انتخاب n واحد از جامعه‌ای به‌حجم N واحد است $(n < N)$ به‌قسمی که همهٔ $k = \binom{N}{n}$ نمونه‌ای که می‌توان انتخاب کرد شانسی برابر برای انتخاب شدن داشته باشند. معمولاً واحدهای جامعه را از ۱ تا N شماره‌گذاری، و سپس به‌طور تصادفی و یا با چشم بسته واحدی را انتخاب می‌کنند، آن‌گاه بدون برگرداندن این واحد به جامعه به‌تصادف واحد دوم را انتخاب می‌کنند و این فرایند را تا انتخاب n واحد نمونه ادامه می‌دهند. نمونهٔ حاصل از این

n واحد را نمونهٔ تصادفی بدون جایگذاری می‌نامند. در عمل از جدول اعداد تصادفی و یا از بستهٔ نرم‌افزاری کامپیوتری مولد اعداد تصادفی برای انتخاب متوالی این n واحد استفاده می‌کنند. ظاهراً به‌نظر می‌رسد که وقتی واحد اول را از بین N واحد انتخاب می‌کنیم، و انتخاب واحد دوم را لاجرم از بین $N - ۱$ واحد باقی‌مانده انجام می‌دهیم، با ادامهٔ کار نهایتاً شانس انتخاب واحدهای نمونه با هم متفاوت‌اند. ولی این احساس شهودی درست نیست. قضیهٔ زیر معرف این مطلب است.

قضیهٔ ۱.۲ در فرایند انتخاب نمونهٔ تصادفی بدون جایگذاری

الف) شانس انتخاب همهٔ $\binom{N}{n}$ نمونهٔ ممکن یکی است.

ب) احتمال قرارگرفتن هر واحد جامعه در نمونه، برابر با $\frac{۱}{N}$ است.

برهان. الف) نمونه‌ای به‌حجم n که واحدهای آن کاملاً مشخص‌اند در نظر بگیرید. در اولین استخراج از جامعه، احتمال انتخاب یک واحد از این مجموعه برابر با $\frac{n}{N}$ است. در استخراج دوم از جامعه، احتمال انتخاب یکی از $n - ۱$ واحد مشخص، پس از حذف واحد منتخب، $\frac{n-۱}{N-۱}$ است و قس‌علی‌هذا. بنابراین احتمال اینکه n واحد مشخص مزبور، در n استخراج از جامعه، برای عضویت در نمونه انتخاب شوند برابر است با

$$\frac{n}{N}\cdot\frac{n-۱}{N-۱}\cdot\frac{n-۲}{N-۲}\cdots\frac{۱}{N-n+۱}=\frac{n!}{N(N-۱)\cdots(N-n+۱)}$$
$$=\frac{n!(N-n)!}{N!}=\frac{۱}{\binom{N}{n}}$$

یعنی با توجه به اینکه تعداد کل نمونه‌های ممکن به‌حجم n از جامعهٔ به‌حجم N برابر با $K = \binom{N}{n}$ است شانس انتخاب هر زیرمجموعهٔ مشخص به‌حجم n برابر با $\frac{۱}{K} = \frac{۱}{\binom{N}{n}}$ است. لذا قسمت (الف) اثبات می‌شود.

ب) تعبیر دیگر (ب) این است که احتمال استخراج هر واحد مشخص از جامعه در هر یک از n استخراج برای عضویت در نمونه برابر با احتمال استخراج آن واحد مشخص در استخراج اول است. درواقع احتمال استخراج یک واحد مشخص در انتخاب rام برابر با $\frac{۱}{N}$ است. نوبت انتخاب r می‌تواند از ۱ تا n باشد. برای اثبات این ویژگی ابتدا توجه می‌کنیم که پیشامد استخراج واحدی مشخص در انتخاب rام، نتیجهٔ رخداد توأم دو پیشامد زیر است:

A: پیشامد اینکه واحد مشخص، در انتخابهای اول، دوم، . . . ، $(r - ۱)$ام استخراج نشود.

B: پیشامد اینکه این واحد فقط در انتخاب rام استخراج شود.

ابتدا $\Pr(A)$ و $\Pr(B)$ را محاسبه می‌کنیم:

محاسبهٔ $\Pr(A)$: احتمال استخراج واحد مشخص در اولین انتخاب $\frac{۱}{N}$، و در نتیجه احتمال عدم استخراج آن در اولین انتخاب $\frac{N-۱}{N} = ۱ - \frac{۱}{N}$ است. احتمال استخراج این واحد در انتخاب دوم به‌شرط آنکه در انتخاب اول استخراج نشود $\frac{۱}{N-۱}$، و در نتیجه احتمال عدم استخراج آن در

انتخاب دوم $\frac{N-۲}{N-۱} = ۱ - \frac{۱}{N-۱}$ است، و به‌همین ترتیب احتمال عدم استخراج این واحد مشخص در انتخاب $(r-۱)$ام به‌شرط اینکه در انتخابهای قبلی استخراج نشده باشد، برابر با $\frac{N-(r-۱)}{N-r+۲}$ است. لذا احتمال پیشامد A چنین است:

$$\Pr(A) = \frac{N-۱}{N} \cdot \frac{N-۲}{N-۱} \cdot \frac{N-۳}{N-۲} \cdot \ldots \cdot \frac{N-r+۱}{N-r+۲} = \frac{N-r+۱}{N}$$

محاسبهٔ $\Pr(B)$: چون واحد مشخص، در $r-۱$ انتخاب قبلی استخراج نشده است به‌وضوح برای انتخاب از بین $N-(r-۱)$ واحد باقی‌مانده شانسی برابر با $\frac{۱}{N-(r-۱)}$ دارد، یعنی

$$\Pr(B) = \frac{۱}{N-r+۱}$$

حال با در نظر گرفتن اینکه استخراج این واحد مشخص در انتخاب rام انجام شده است، احتمال انتخاب آن برابر با احتمال رخداد توأم دو پیشامد مستقل A و B است. پس احتمال استخراج واحدی مشخص در انتخاب rام برابر است با

$$\Pr(A) \cdot \Pr(B) = \frac{N-r+۱}{N} \cdot \frac{۱}{N-r+۱} = \frac{۱}{N}$$

پس، نتیجه می‌گیریم احتمال اینکه واحدی مشخص در نمونه‌ای به‌حجم n و در استخراجی معین ظاهر شود، برابر با $\frac{۱}{N}$ است و قسمت (ب)ی قضیهٔ ثابت می‌شود. □

گاهی ویژگی (ب) را به‌صورت نادرست برای تعریف نمونهٔ تصادفی سادهٔ بدون جایگذاری به‌کار می‌برند بدون اینکه توجه کنند روشهای نمونه‌گیری دیگری هم وجود دارند که دارای همین ویژگی هستند. در مثال زیر ویژگی (ب) را نشان داده‌ایم.

مثال ۱.۲ فرض می‌کنیم جامعه از چهار واحد $u = \{۱, ۲, ۳, ۴\}$ تشکیل شده باشد. نمونه‌ای به حجم ۳ به‌روش تصادفی بدون جایگذاری انتخاب می‌کنیم. نشان می‌دهیم که احتمال استخراج مثلاً واحد مشخص ۲ در هر یک از انتخابهای اول و یا دوم و یا سوم، برابر با $\frac{۱}{۴}$ است.

می‌دانیم تعداد نمونه‌های به‌حجم ۳ از جامعهٔ به‌حجم ۴ برابر $۴ = \binom{۴}{۳}$ است. نمونه‌های به‌حجم ۳ عبارت‌اند از

$$(۱, ۲, ۳) \quad (۱, ۲, ۴) \quad (۱, ۳, ۴) \quad (۲, ۳, ۴)$$

چون در این مثال جای ۲ که نشان‌دهندهٔ شمارهٔ انتخاب است نیز اهمیت دارد، هر نمونه می‌تواند به $۶ = ۳!$ صورت ظاهر شود. لذا کل صورتهای انتخاب نمونه‌ها چنین‌اند:

(۱, ۲, ۳)	(۱, ۲, ۴)	(۱, ۳, ۴)	(۲, ۳, ۴)
(۱, ۳, ۲)	(۱, ۴, ۲)	(۱, ۴, ۳)	(۲, ۴, ۳)
(۲, ۱, ۳)	(۲, ۱, ۴)	(۳, ۱, ۴)	(۳, ۲, ۴)
(۲, ۳, ۱)	(۲, ۴, ۱)	(۳, ۴, ۱)	(۳, ۴, ۲)
(۳, ۱, ۲)	(۴, ۱, ۲)	(۴, ۱, ۳)	(۴, ۲, ۳)
(۳, ۲, ۱)	(۴, ۲, ۱)	(۴, ۳, ۱)	(۴, ۳, ۲)

در صورتهای بالا به‌سهولت می‌توان دید که ۲ به‌عنوان انتخاب دوم در ۶ نمونه ظاهر شده است لذا احتمال استخراج آن در دومین انتخاب $\frac{۱}{۴} = \frac{۶}{۲۴}$ است، زیرا کل تعداد نمونه‌ها با توجه به‌ترتیب انتخاب، ۲۴ تاست. به‌همین ترتیب، ۲ در انتخاب اول در ۶ نمونه، در انتخاب سوم و چهارم هم، هرکدام در ۶ نمونه ظاهر شده است و لذا احتمال استخراج آن در هر انتخاب مقدار ثابت $\frac{۱}{۴}$ است. این مطلب برای ۱، ۳، و ۴ نیز صحیح است. ▲

۲.۲ انتخاب نمونهٔ تصادفی سادهٔ بدون جایگذاری

جدول اعداد تصادفی، متشکل از برآمدهای امتحانهای تصادفی مستقل از توزیع احتمال یکنواخت گسسته است. هر امتحان دارای برآمدهای ممکن ۰، ۱، ۲، . . . ، ۹ با احتمال برابر است. لذا در هر خانهٔ جدول، تمام ارقام ۰ تا ۹ برای ظاهر شدن در آن خانه، احتمالی یکسان دارند و برای خانه‌های مختلف جدول برآمدها مستقل‌اند. ساختار جدول به‌گونه‌ای است که اگر اعداد ۲ یا ۳ یا . . .رقمی انتخاب کنیم باز هم احتمال انتخاب، ثابت باقی می‌ماند.

جدولهای موجود اعداد تصادفی به‌وسیلهٔ نرم‌افزارهای کامپیوتری تولید می‌شوند. شرکت راند جدولی با یک میلیون رقم منتشر کرده است. جداول متعدد دیگری نیز موجودند. جدول I شامل ۲۰۰۰ رقم تصادفی است که برای استفاده در مثالها و مسائل این کتاب کافی است. وقتی حجم نمونه بزرگ باشد استفاده از بسته‌های کامپیوتری راحت‌تر از استفاده از جدول ارقام تصادفی است. به‌ویژه اغلب این بسته‌ها اعداد تصادفیی را که انتخاب می‌کنند مرتب نیز می‌نمایند. استفاده‌کننده معمولاً با توجه به حجم جامعه، کوچکترین و بزرگترین عدد تصادفی مورد نیاز و حجم نمونه را به برنامه می‌دهد و اعداد تصادفی منتخب مرتب را در خروجی دریافت می‌کند. برای اینکه استفاده از جدول اعداد تصادفی را تمرین کنید مثالی را مطرح می‌کنیم.

مثال ۲.۲ می‌خواهیم از ۴۲۷ صورتحساب که از ۱ تا ۴۲۷ شماره‌گذاری شده‌اند نمونه‌ای به‌حجم ۱۲ انتخاب کنیم و درستی آنها را بررسی نماییم.

در استفاده از جدول اعداد تصادفی ابتدا باید واحدهای جامعه را شماره‌گذاری کرد. معمولاً این شماره‌گذاری از ۱ تا N که حجم جامعه است صورت می‌گیرد ولی در صورت لزوم می‌توان شماره‌ها را از هر عدد a تا $a + N$ در نظر گرفت. مثلاً در مثال مورد بحث ممکن است خود صورتحسابها شماره‌های متوالی از ۱۰۱ تا ۵۲۷ داشته باشند. در این صورت ضرورت ندارد که

شماره‌ها را از ۱ شروع کنیم. در این مثال باید از شماره‌های ۱ تا ۴۲۷، ۱۲ عدد به‌تصادف انتخاب کنیم تا شمارهٔ واحدهای نمونه، یعنی صورتحسابهای مربوط به نمونه مشخص شوند. چون N سه‌رقمی است سه ستون متوالی از ستونهای یکی از صفحات جدول I، مثلاً ستونهای ۱۰ تا ۱۲ صفحهٔ ۴۵ را انتخاب و از بالا به پایین در امتداد ۳ ستون حرکت می‌کنیم و ۱۲ عدد را که از ۰۰۰ تا ۴۲۶ هستند یادداشت می‌نماییم. اگر عدد ۰۰۰ جزء اعداد منتخب بود آن را معادل عدد ۴۲۷ می‌گیریم. در بالا، دو عد ۷۸۲، ۴۷۳ را نادیده می‌گیریم. اولین عدد مطلوب ۱۳۳ است. به‌تدریج که به‌سوی پایین می‌رویم به‌ترتیب اعداد زیر نتیجه می‌شوند:

۱۳۳ ۳۲۵ ۲۶۷ ۳۱۹ ۱۳۴ ۰۵۵ ۰۹۲ ۳۳۰ ۰۷۱

۳۳۸ ۴۱۵ ۳۵۸

بدیهی است در حرکت به‌سوی پایین، بسیاری از اعداد سه‌رقمی، به‌دلیل اینکه بزرگتر از ۴۲۷ هستند یادداشت نشده‌اند. دوازده عدد بالا شماره‌های صورتحسابهایی هستند که باید برای بررسی، واحدهای نمونه باشند.
ۀ

اگر بخواهیم چندین نمونهٔ تصادفی از جامعه استخراج کنیم باید نقاط شروع خود در جدول را تغییر دهیم. در این روش به‌دلیل حذف بسیاری از اعداد، مسلماً هزینه و وقتی اضافی صرف می‌شود. گاهی روش دیگری را که با حذف کمتر اعداد تصادفی همراه است به‌کار می‌برند. برای شرح دادن این روش، فرض کنید $N = ۱۵۶$. سه ستون ۱۷، ۱۸، ۱۹ صفحهٔ ۴۵ در نظر می‌گیریم. اگر اعداد سه‌رقمی این ستونها بین ۲۰۰ و ۴۰۰ باشند ۲۰۰ تا از آنها کم می‌کنیم و اگر بین ۴۰۰ و ۶۰۰ باشند، ۴۰۰ تا از آنها می‌کاهیم و قس‌علی‌هذا و نتایج را اگر از ۱ تا ۱۵۶ باشند به‌عنوان شمارهٔ واحدهایی از جامعه که باید عضو نمونه باشند یادداشت می‌کنیم. مثلاً، اگر $n = ۸$، اعداد زیر از ستونهای بالا نتیجه می‌شوند:

۱۰۸ ۵۱ ۰۳۸ ۰۲ ۱۴۹ ۱۴۲ ۹۴ ۴۱

در عمل ملاحظه می‌شود که با این خط‌مشی، تعداد اعدادی که حذف می‌شوند اندک است. گاهی برای صرفه‌جویی در وقت، می‌توان اگر اعدادی بزرگتر از N انتخاب شوند با روش باقیمانده تقسیم آنها بر N از این اعداد استفاده کرد. مثلاً اگر در مثال بالا عدد ۵۲۸ انتخاب شود با توجه به‌تقسیم $۵۲۸ \div ۴۲۷ = ۱ + \frac{۱۰۱}{۴۲۷}$ عدد ۱۰۱ را که باقی‌مانده است انتخاب می‌کنیم.

۳.۲ تعریفها و نمادها

در بررسی نمونه‌ای، توجه ما به بعضی ویژگیهای واحدهاست و هدف این است که به‌کمک نمونه، پارامترهایی از جامعه را که در رابطه با این ویژگیها هستند برآورد کنیم. هر ویژگی مورد توجه را یک

جدول ۱ اعداد تصادفی

67 00	76 07	06 04	17 26	85 10	29 42	93 48	93 46	52 72	77 53
37 41	48 98	99 14	86 78	56 14	20 12	28 80	70 70	66 62	99 86
54 85	60 58	43 58	30 74	44 33	96 38	13 52	98 74	01 27	62 08
82 78	21 26	47 21	31 06	50 67	34 87	78 86	26 32	35 38	94 63
72 32	72 25	83 98	34 31	63 44	31 47	09 57	26 23	89 88	16 10
86 73	37 38	09 68	16 67	81 32	03 42	28 56	09 92	75 20	50 35
54 67	40 72	87 91	00 61	98 95	38 02	94 57	65 32	73 34	64 33
80 86	35 17	08 51	17 12	07 87	75 39	83 43	77 04	66 02	13 46
08 32	44 20	01 13	17 22	42 71	76 76	33 56	94 22	02 67	70 98
96 84	83 43	36 80	18 75	16 54	53 48	71 77	34 88	43 61	41 76
43 67	84 20	48 23	50 47	15 85	24 65	78 93	01 84	02 04	41 31
35 99	47 15	37 62	62 27	35 41	55 57	03 12	74 45	83 25	14 67
13 07	22 58	68 80	91 93	64 68	59 55	18 45	72 83	08 01	28 93
73 15	83 78	76 46	78 36	65 56	34 75	92 58	99 38	51 64	98 42
18 92	29 56	47 99	74 31	42 88	52 71	90 84	23 56	75 22	62 08
60 07	11 21	26 62	94 01	89 32	61 14	17 11	30 31	12 01	18 58
59 60	63 71	99 35	15 56	41 95	71 78	53 15	10 51	80 17	53 81
45 55	85 24	55 08	49 43	00 21	31 67	75 35	42 10	71 12	46 37
90 80	65 04	38 06	30 57	56 62	21 88	30 86	56 80	02 21	43 40
84 51	93 90	28 31	22 31	48 44	45 97	48 85	79 68	78 78	05 18
07 66	01 78	75 25	68 67	31 08	85 38	37 76	01 94	22 20	03 04
19 41	96 21	21 48	63 68	46 91	11 40	98 12	50 26	58 62	74 30
01 38	53 01	20 30	43 53	83 34	87 15	63 52	17 89	43 19	31 11
12 95	21 94	99 72	76 51	69 20	66 93	80 83	88 97	35 52	23 76
25 88	63 69	99 41	89 27	18 92	52 49	56 75	99 20	68 13	04 50
95 89	07 45	38 96	63 61	11 49	98 72	60 67	30 94	93 01	20 20
49 69	36 31	40 43	65 22	63 59	43 94	43 18	76 48	00 90	10 65
47 52	59 03	71 19	04 67	42 38	98 78	36 75	12 62	19 27	23 83
41 89	34 25	98 99	14 49	65 61	20 09	71 32	63 20	88 92	25 40
41 89	18 07	02 57	18 44	53 64	89 51	56 63	63 37	25 64	17 23
46 58	12 07	61 94	29 39	90 76	24 23	64 83	38 61	35 84	78 95
98 42	17 61	53 32	62 34	19 38	05 03	07 09	45 01	61 01	81 34
09 44	61 42	84 40	80 09	25 36	73 61	09 53	51 95	76 09	13 64
41 97	74 05	94 04	57 50	28 49	26 54	91 60	26 20	76 12	91 39
70 42	82 33	21 08	41 30	67 58	46 56	84 19	40 76	47 37	85 59
05 18	96 66	53 07	84 44	17 62	70 43	76 28	64 80	98 32	21 11
69 44	33 07	09 02	87 76	98 60	65 99	36 27	77 23	93 92	15 72
71 95	73 70	09 66	69 66	73 19	20 69	12 95	01 99	75 83	31 13
99 59	52 07	54 56	90 44	76 85	84 35	17 08	97 87	56 04	61 52
97 07	78 13	46 90	10 48	53 29	43 92	58 61	39 39	18 38	47 35
85 04	86 52	92 49	65 46	99 78	99 66	82 34	22 86	79 10	85 06
11 68	36 63	15 84	92 56	31 78	47 49	14 61	34 78	76 47	84 47
12 69	35 64	97 00	63 69	41 06	75 10	94 21	70 74	00 08	90 56
62 72	73 46	26 19	35 75	15 23	76 26	98 66	97 45	31 86	44 80
78 63	02 76	61 95	67 00	30 05	18 62	19 86	40 08	83 32	17 42
65 40	31 04	87 02	46 38	43 16	63 83	76 95	23 06	78 48	54 60
42 68	22 96	29 30	39 32	75 36	64 03	70 64	83 51	81 81	16 96
40 15	54 28	80 30	30 07	53 01	62 62	26 31	75 25	10 23	43 84
51 19	95 91	95 98	92 53	98 08	55 70	68 78	21 13	95 15	87 36
77 56	25 60	17 30	63 23	98 29	62 71	92 10	71 72	52 21	06 21

جدول I (ادامه)

65 23	68 00	77 82	58 14	10 85	11 85	57 11	73 74	45 25	60 46
06 56	76 51	04 73	94 30	16 74	69 59	04 38	83 98	30 20	87 85
55 99	98 60	01 33	06 93	85 13	23 17	26 51	92 04	52 31	38 70
72 82	45 44	09 53	04 83	03 83	98 41	67 41	01 38	66 83	11 99
04 21	28 72	73 25	02 74	35 81	78 49	52 67	61 40	60 50	47 50
87 01	80 59	89 36	41 59	60 27	64 89	47 45	18 21	69 84	76 06
31 62	46 53	84 40	56 31	74 96	52 23	72 95	96 06	56 83	85 22
29 81	57 94	35 91	90 70	94 24	19 35	50 22	23 72	87 34	83 15
39 98	74 22	77 19	12 81	29 42	04 50	62 34	36 81	43 07	97 92
56 14	80 10	76 52	38 54	84 13	99 90	22 55	41 04	72 37	89 33
29 56	62 74	12 67	09 35	89 33	04 28	44 75	01 57	87 45	52 21
93 32	57 38	39 36	87 42	72 55	73 97	98 36	57 41	76 09	11 68
95 69	51 54	43 19	20 49	57 25	90 55	26 20	70 98	43 73	56 45
65 71	32 43	64 67	22 55	65 65	48 86	10 88	20 12	40 18	49 25
90 27	33 43	97 84	20 57	49 91	41 20	17 64	29 60	66 87	55 97
95 29	42 45	61 34	30 13	30 39	21 52	59 28	64 98	08 76	09 27
99 74	06 29	20 55	72 70	11 43	95 82	75 37	90 24	77 43	63 21
87 87	56 91	16 97	51 50	61 36	96 47	76 68	49 11	50 56	51 06
46 24	17 74	97 37	39 03	54 83	34 00	74 61	77 51	43 63	15 67
66 79	81 43	40 92	84 72	88 32	83 24	67 01	41 34	70 19	26 93
36 42	94 58	83 30	92 39	18 40	03 00	12 90	32 27	91 65	48 15
07 66	25 08	99 27	69 48	85 32	16 46	19 31	85 02	86 36	22 96
93 10	05 72	18 26	36 67	68 48	31 69	68 58	93 49	45 86	99 29
49 50	63 99	26 71	47 94	32 71	72 91	34 18	74 06	32 14	40 80
20 75	58 89	39 04	42 73	37 93	11 07	28 77	91 36	60 47	82 62
02 40	62 09	00 71	09 37	80 44	50 37	32 70	20 38	71 86	75 34
59 87	21 38	29 78	72 67	42 83	65 21	54 79	66 42	47 86	31 15
48 08	99 66	43 38	28 13	50 25	47 93	11 15	07 84	28 30	19 07
54 26	86 75	44 15	20 39	20 03	58 54	80 29	62 53	06 97	71 51
35 35	58 45	23 58	63 66	09 62	80 92	14 55	81 41	21 48	87 34
73 84	90 49	01 21	90 29	57 06	68 73	51 10	51 95	63 08	57 99
34 64	78 00	92 59	67 74	58 48	92 09	42 20	40 37	63 80	58 93
68 56	87 47	63 06	24 71	41 98	79 06	07 18	58 29	16 49	67 37
72 47	05 42	88 07	27 55	58 74	82 08	42 28	26 48	25 32	00 31
44 44	96 75	89 57	12 60	42 38	77 36	45 69	21 68	32 70	04 96
28 11	57 47	61 57	89 88	62 18	93 67	57 32	96 72	21 17	13 54
87 22	38 88	91 99	16 08	17 76	27 47	52 14	98 86	35 68	23 85
44 93	14 59	67 40	24 10	11 63	40 47	07 56	14 22	62 74	93 39
81 84	37 25	90 43	56 62	94 58	49 03	84 22	57 22	47 98	86 37
09 75	35 21	04 47	54 08	98 44	08 16	44 86	69 71	20 52	64 94
77 65	05 04	22 18	20 10	81 87	05 69	43 70	96 76	42 05	21 10
19 06	51 61	34 03	61 55	98 58	83 50	01 48	99 85	08 67	15 91
52 91	87 07	19 62	32 28	04 91	42 48	65 24	86 09	87 68	55 51
52 47	25 14	93 91	75 51	49 26	49 41	20 83	30 30	43 22	69 08
52 67	87 40	63 41	91 86	10 47	80 70	56 87	25 86	89 94	21 42
65 25	71 73	78 60	50 62	91 04	95 97	64 16	71 31	32 80	19 61
29 97	56 42	56 90	16 75	74 95	99 26	01 63	25 16	54 18	54 46
15 25	03 68	92 45	53 00	06 29	46 43	46 66	27 12	85 05	22 44
82 08	65 67	64 13	51 14	38 28	24 30	39 62	20 35	23 90	57 36
81 35	03 25	87 24	83 59	04 67	51 52	26 21	69 75	87 28	61 50

مشخصه یا صفت می‌نامند. حجم جامعه را با N نشان می‌دهیم و اندازهٔ مشخصهٔ تحت بررسی واحدهای جامعه را به‌صورت

$$Y_۱ \quad Y_۲ \quad Y_۳ \cdots Y_N$$

نمایش می‌دهیم. وقتی به‌کمک جدول اعداد تصادفی یا بستهٔ کامپیوتری، نمونه‌ای به‌حجم n از این جامعه تهیه می‌کنیم اندازهٔ مشخصهٔ تحت بررسی واحدهای نمونه را به‌صورت

$$Y_۱ \quad Y_۲ \quad Y_۳ \cdots Y_n$$

نشان می‌دهیم. غالباً $Y_۱$ در نمونه با $Y_۱$ در جامعه متفاوت است. ممکن است $Y_۱$ نمونه یعنی اولین واحدی که به‌تصادف برای عضویت در نمونه انتخاب می‌شود مثلاً $Y_{۲۰}$ جامعه باشد. به‌همین ترتیب امکان دارد $Y_۲$ی نمونه با $Y_۲$ی جامعه متفاوت باشد و قس‌علی‌هذا. نمایش نوعی اندازه‌های واحدهای جامعه به‌صورت، $(i = ۱, \cdots, N)Y_i$ و نمایش نوعی اندازه‌های واحدهای نمونه به‌همان صورت $(i = ۱, \cdots, n)Y_i$ است. میانگین جامعه را با $\overline{Y}_N$ و میانگین نمونه را با $\overline{Y}_n$ نشان می‌دهیم، پس

$$\overline{Y}_N = \frac{۱}{N}\sum_{i=۱}^{N} Y_i \quad , \quad \overline{Y}_n = \frac{۱}{n}\sum_{i=۱}^{n} Y_i$$

در مورد یک نمونهٔ بخصوص، $\overline{Y}_n$ عددی مشخص است ولی از لحاظ نظری چون تعداد نمونه‌های تصادفی بدون جایگذاری $\binom{N}{n}$ تاست محققاً $K = \binom{N}{n}$ تا $\overline{Y}_n$ داریم که ممکن است برخی از آنها با هم برابر باشند. لذا می‌توان $\overline{Y}_n$ را متغیری تصادفی تلقی کرد که در عمل تنها یکی از مقادیرش رخ می‌دهد. میانگین یک نمونهٔ مشخص را هم با $\overline{Y}_n$ نشان می‌دهیم. خواننده معمولاً می‌تواند در متن هر مطلبی تشخیص دهد که هدف از نماد $\overline{Y}_n$ چه موقع میانگین یک نمونهٔ مشخص، یعنی درواقع یک عدد است و چه موقع معرّف متغیری تصادفی است که همهٔ مقادیر ممکن میانگین نمونه‌ها را می‌پذیرد. میانگین جامعه، $\overline{Y}_N$، عددی ثابت ولی مجهول است و سعی ما بر این است که آن را به‌کمک نمونه برآورد کنیم. مطابق معمول برآوردکنندهٔ $\overline{Y}_N$ را با $\hat{\overline{Y}}_N$ نشان می‌دهیم. مجموع مقادیر مشخصهٔ واحدهای جامعه، یعنی $\sum_{i=۱}^{N} Y_i$ را با T_N و مجموع مقادیر مشخصهٔ واحدهای نمونه به‌حجم n را با T_n نشان می‌دهیم. بدیهی است T_N پارامتری ثابت ولی مجهول است، در حالی که T_n یک آماره، یعنی متغیری تصادفی و برای نمونه‌ای مشخص، عددی معلوم است. برآوردکنندهٔ T_N را طبق معمول با $\hat{T}_N$ نمایش می‌دهیم. از این به‌بعد برای سادگی در نوشتن، به‌جای مجموع مقادیر مشخصهٔ واحدهای جامعه یا نمونه، از اصطلاح «مجموع واحدهای جامعه» یا «مجموع واحدهای نمونه» استفاده می‌کنیم. نمادهای بالا را در جدول زیر خلاصه کرده‌ایم، پس

جدول ۱.۲ معرفی نمادهای مورد استفاده در نمونه‌گیری

$\overline{Y}_N$	پارامتر: میانگین جامعه	مقدار ثابت	برآوردکنندهٔ $\overline{Y}_N$: $\hat{\overline{Y}}_N$
$\overline{Y}_n$	آماره: میانگین نمونه	متغیر تصادفی	برای یک نمونهٔ مشخص، $\overline{Y}_n$ مقداری است ثابت
T_N	پارامتر: مجموع واحدهای جامعه	مقدار ثابت	برآوردکنندهٔ T_N : $\hat{T}_N$
T_n	آماره: مجموع واحدهای نمونه	متغیر تصادفی	برای یک نمونهٔ مشخص، T_n مقداری است ثابت

از ذکر این مطالب و نمادها سعی می‌کنیم از روی نمونهٔ تصادفی بدون جایگذاری، برآوردکننده‌ای نااریب برای $\overline{Y}_N$ به‌دست آوریم.

قضیهٔ ۲.۲ میانگین نمونهٔ تصادفی سادهٔ بدون جایگذاری، برآوردکنندهٔ نااریب میانگین جامعه است.

برهان. اگر حجم نمونهٔ تصادفی بدون جایگذاری را n و حجم جامعه را N بگیریم، باید ثابت کنیم که

$$E(\overline{Y}_n) = \overline{Y}_N \qquad (۱.۲)$$

همان‌طور که می‌دانیم تعداد نمونه‌های تصادفی بدون جایگذاری ممکن، $K = \binom{N}{n}$ است. لذا $K = \binom{N}{n}$ تا $\overline{Y}_n$ داریم که مقادیر متغیر تصادفی $\overline{Y}_n$ هستند. اگر قرار دهیم $K = \binom{N}{n}$، داریم

$$E(\overline{Y}_n) = \frac{\sum \overline{Y}_n}{K}$$

که در آن، $\sum \overline{Y}_n$ مجموع همهٔ K مقداری است که متغیر تصادفی $\overline{Y}_n$ می‌پذیرد. برای یک نمونهٔ نوعی

$$\overline{Y}_n = \frac{۱}{n}(Y_۱ + Y_۲ + \cdots + Y_n)$$

پس

$$E(\overline{Y}_n) = \frac{\sum(Y_۱ + Y_۲ + \cdots + Y_n)}{nK} \qquad (۲.۲)$$

منظور از $\sum(Y_۱ + Y_۲ + \cdots + Y_n)$ مجموع عناصر تمام نمونه‌های nتایی ممکن است. وقتی همهٔ عناصر نمونه‌های ممکن را با هم جمع کنیم، مایلیم بدانیم در این مجموع هر واحد جامعه چندبار ظاهر می‌شود. دانستن این مطلب، کار مجموعیابی را تسهیل می‌کند.

واحد Y_iی جامعه را در نظر بگیرید. درک اینکه Y_i در چندتا از K نمونهٔ ممکن ظاهر می‌شود آسان است. بدین طریق که Y_i را از جامعه کنار می‌گذاریم و همهٔ نمونه‌های تصادفی سادهٔ به‌حجم $n-۱$ حاصل از جامعهٔ به‌حجم $N-۱$ را در نظر می‌گیریم. تعداد این نمونه‌ها $\binom{N-۱}{n-۱}$ تاست. حال اگر به هر یک از این نمونه‌ها عنصر کنارگذاشته‌شدهٔ Y_i را ضمیمه کنیم همهٔ نمونه‌های با حجم n که از جامعهٔ به‌حجم N استخراج شده‌اند، و در عین حال شامل Y_i هستند به‌دست می‌آیند. پس تعداد نمونه‌های nتایی شامل هر واحد Y_i، برابر با $\binom{N-۱}{n-۱}$ است.

به‌عبارت دیگر درکل نمونه‌های ممکن nتایی جامعه، هر واحد $\binom{N-۱}{n-۱}$ بار ظاهر می‌شود. پس در $\sum(Y_۱+Y_۲+\cdots+Y_n)$ که صورت کسر طرف دوم (۲.۲) است، هر واحد جامعه به تعداد $\binom{N-۱}{n-۱}$ بار ظاهر می‌شود. پس اگر T_N مجموع واحدهای جامعه باشد

$$\sum(Y_۱+Y_۲+\cdots+Y_n)=\binom{N-۱}{n-۱}T_N$$

از طرفی

$$\overline{Y}_N=\frac{T_N}{N}$$

پس

$$\sum(Y_۱+Y_۲+\cdots+Y_n)=\binom{N-۱}{n-۱}N\overline{Y}_N$$

لذا (۲.۲) به‌صورت زیر در می‌آید

$$\begin{aligned}E(\overline{Y}_n)&=\frac{\binom{N-۱}{n-۱}N\overline{Y}_N}{nK}=\frac{N\binom{N-۱}{n-۱}}{n\binom{N}{n}}\cdot\overline{Y}_N\\&=\frac{N}{n}\frac{(N-۱)!/\{(n-۱)!(N-n)!\}}{N!/\{n!(N-n!)\}}\cdot\overline{Y}_N=\frac{N}{n}\cdot\frac{n}{N}\overline{Y}_N=\overline{Y}_N\end{aligned}$$

و حکم قضیه ثابت می‌شود. این برابری حاکی از آن است که اگر میانگین همهٔ میانگینهای K نمونهٔ ممکن را بیابیم میانگین واقعی جامعه به‌دست می‌آید. در عمل تنها یکی از نمونه‌های ممکن، و در نتیجه تنها یکی از K میانگین $\overline{Y}_n$، یعنی یکی از مقادیر متغیر تصادفی $\overline{Y}_n$ را داریم. این مقدار برآورد ناااریب میانگین جامعه است که در حالت کلی با $\overline{Y}_N$ واقعی متفاوت است. □

مثال ۳.۲ جامعه‌ای شامل ۵ واحد است:

$$\{۳,\ ۵,\ ۶,\ ۴,\ ۲\}$$

نمونه‌ای به‌حجم ۴ از آن به‌تصادف و بدون جایگذاری انتخاب می‌کنیم، نتیجهٔ کار یکی از $۵ = \binom{۵}{۴}$ نمونهٔ ممکن زیر است

$$(۳,\ ۵,\ ۶,\ ۴)$$
$$(۲,\ ۵,\ ۶,\ ۴)$$
$$(۳,\ ۲,\ ۶,\ ۴)$$
$$(۳,\ ۵,\ ۲,\ ۴)$$
$$(۳,\ ۵,\ ۶,\ ۲)$$

میانگین این ۵ نمونه به‌ترتیب عبارت‌اند از

$$\overline{Y}_n : \frac{۱۸}{۴},\ \frac{۱۷}{۴},\ \frac{۱۵}{۴},\ \frac{۱۴}{۴},\ \frac{۱۶}{۴}$$

اگر میانگین این پنج میانگین ممکن را حساب کنیم، نتیجه می‌شود

$$E(\overline{Y}_n) = \frac{۱}{۵} \cdot \frac{۸۰}{۴} = ۴$$

می‌بینید که اگر میانگین جامعه را حساب کنیم $\overline{Y}_N$ برابر با ۴ به‌دست می‌آید. پس برابری $E(\overline{Y}_n) = \overline{Y}_N = ۴$ با حکم قضیه مطابقت دارد. همان‌طورکه متذکر شدیم در عمل تنها یکی از پنج نمونه دردست است. مثلاً فرض کنید نمونهٔ تصادفی منتخب، اولین نمونه باشد. در این صورت میانگین این نمونه، یعنی $۴٫۵ = \frac{۱۸}{۴}$ رویدادی از متغیر تصادفی $\overline{Y}_n$ است و برآورد ناااریب میانگین جامعه خواهد بود. به‌عبارت دیگر اگر ندانیم که میانگین جامعه برابر با ۴ است، با مشاهدهٔ نمونهٔ اول، ۴٫۵ را به‌عنوان برآورد جامعه در نظر می‌گیریم. ▲

۲.۴ واریانس میانگین نمونه

در بخش قبل دیدیم که در عمل یکی از K مقدار ممکن $\overline{Y}_n$ را به‌کمک مشاهدهٔ یک تک نمونه به‌دست می‌آوریم و آن را به عنوان برآورد نااریب $\overline{Y}_N$ انتخاب می‌کنیم. ممکن است $\overline{Y}_n$ از $\overline{Y}_N$ واقعی بسیار دور و یا به آن نزدیک باشد. مسلماً نزدیکی $\overline{Y}_n$ به $\overline{Y}_N$ واقعی، حاکی از دقت نتیجه است. اگر دنبالهٔ K مقدار $\overline{Y}_n$ را در نظر بگیرید جمله‌های دنباله، حول $\overline{Y}_N$ واقعی که میانگین این جمله‌هاست پراکنده‌اند. اگر این جمله‌ها حول $\overline{Y}_N$ و بسیار نزدیک به آن پراکنده باشند، وقتی در عمل یکی از آنها رخ می‌دهد این یکی هم به $\overline{Y}_N$ نزدیک خواهد بود و برآوردی دقیق برای $\overline{Y}_N$ است. اگر این جمله‌ها حول $\overline{Y}_N$ ولی با فاصله‌های زیاد پراکنده باشند ممکن است $\overline{Y}_n$ حاصل از نمونهٔ مشاهده‌شده، مقداری فرین باشد، در نتیجه ضمن اینکه برآورد نااریب $\overline{Y}_N$ است، ولی با

آن فاصله زیاد دارد و برآوردی نااریب ولی دور از واقعیت است. نتیجهٔ بحث این است که وقتی جمله‌های دنبالهٔ $\overline{Y}_n$ها نزدیک $\overline{Y}_N$ پراکنده باشند، یعنی وقتی واریانس $\overline{Y}_n$ کوچک باشد، مقدار حاصل از نمونهٔ مشاهده‌شده، برآورد نااریب دقیقتری برای $\overline{Y}_N$ است. لذا باید برای بررسی دقت برآوردکنندهٔ $\overline{Y}_n$، واریانس آن را بررسی کنیم. هرچه $V(\overline{Y}_n)$ کوچکتر باشد دقت برآوردکننده بیشتر است. برای محاسبهٔ واریانس $\overline{Y}_n$ به قضیهٔ زیر نیاز داریم:

قضیهٔ ۳.۲ اگر Y_i و Y_j دو واحد مشخص نمونهٔ تصادفی ساده بدون جایگذاری در دو انتخاب متوالی باشند، آن‌گاه

$$\operatorname{cov}(Y_i,\ Y_j) = \frac{-\sigma^۲}{N-۱} \tag{۳.۲}$$

که در آن، $\sigma^۲$ واریانس جامعهٔ به‌حجم N است.

برهان. می‌دانیم

$$\overline{Y}_N = \frac{۱}{N}(Y_۱ + Y_۲ + \cdots + Y_N)$$

پس

$$Y_۱ + Y_۲ + \cdots + Y_N - N\overline{Y}_N = ۰$$

لذا

$$(Y_۱ + Y_۲ + \cdots + Y_N - N\overline{Y}_N)^۲ = ۰$$

یا

$$[(Y_۱ - \overline{Y}_N) + (Y_۲ - \overline{Y}_N) + \cdots + (Y_N - \overline{Y}_N)]^۲ = ۰$$

در نتیجه

$$\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲ + \sum_{i=۱}^{N}\sum_{\substack{j=۱ \\ i\neq j}}^{N}(Y_i - \overline{Y}_N)(Y_j - \overline{Y}_N) = ۰$$

چون $\sigma^۲ = \frac{۱}{N}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲$، پس

$$\sum_{i=۱}^{N}\sum_{\substack{j=۱ \\ i\neq j}}^{N}(Y_i - \overline{Y}_N)(Y_j - \overline{Y}_N) = -N\sigma^۲ \tag{الف}$$

بدیهی است که احتمال انتخاب Y_i از بین N واحد جامعه برابر $\frac{1}{N}$ و احتمال انتخاب واحد Y_j از بین $N-1$ واحد باقیمانده برابر با $\frac{1}{N-1}$ است. لذا احتمال انتخاب متوالی زوج (Y_i, Y_j) برابر با $\frac{1}{N(N-1)}$ است. بنابراین با توجه به تعریف کوواریانس، داریم

$$\operatorname{cov}(Y_i, Y_j) = \sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}\frac{1}{N(N-1)}(Y_i - \overline{Y}_N)(Y_j - \overline{Y}_N)$$

اگر (الف) را در این رابطه منظور کنیم، به‌دست می‌آوریم

$$\operatorname{cov}(Y_i, Y_j) = -N\sigma^2 \cdot \frac{1}{N(N-1)} = \frac{-\sigma^2}{N-1}$$

که برهان را کامل می‌کند. □

اینک با استفاده از قضیهٔ ۳.۲ به‌محاسبهٔ واریانس $\overline{Y}_n$ می‌پردازیم.

قضیهٔ ۴.۲ اگر $\overline{Y}_n$، میانگین نمونهٔ تصادفی بدون جایگذاری با حجم n از جامعه‌ای به‌حجم N باشد، آن‌گاه

$$\mathrm{V}(\overline{Y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 = (1-f)\frac{S^2}{n}$$

که در آن S^2، تغییرات جامعه به‌صورت

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y}_N)^2$$

و $f = \frac{n}{N}$ موسوم به کسر نمونه‌گیری است.

برهان. با توجه به $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$، داریم

$$\mathrm{V}(\overline{Y}_n) = \mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\mathrm{V}(Y_1 + Y_2 + \cdots + Y_n)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{V}(Y_i) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\operatorname{cov}(Y_i, Y_j)\right]$$

هر Y_i، به‌عنوان متغیری تصادفی که می‌تواند هر واحد جامعه باشد، واریانسی دارد که برابر با واریانس جامعه، یعنی σ^2 است. ضمناً با توجه به قضیهٔ ۳.۲، $\operatorname{cov}(Y_i, Y_j) = -\frac{\sigma^2}{N-1}$، پس

رابطهٔ بالا به‌صورت زیر درمی‌آید

$$
\begin{aligned}
\mathrm{V}(\overline{Y}_n) &= \frac{۱}{n^۲}\left[\sum_{i=۱}^{n}\sigma^۲ + \sum_{\substack{i=۱\\ i\neq j}}^{n}\sum_{j=۱}^{n}\left(\frac{-\sigma^۲}{N-۱}\right)\right.\\
&= \frac{۱}{n^۲}\left[n\sigma^۲ + n(n-۱)\left(-\frac{\sigma^۲}{N-۱}\right)\right]\\
&= \frac{\sigma^۲}{n}\left[۱-\frac{n-۱}{N-۱}\right] = \frac{\sigma^۲}{n}\cdot\frac{N-n}{N-۱} \qquad (۴.۲)
\end{aligned}
$$

فرمول بالا واریانس $\overline{Y}_n$ را برحسب واریانس جامعه بیان می‌کند. اما می‌دانیم که $(N-۱)S^۲ = N\sigma^۲ = \sum_{i=۱}^{N}(Y_i-\overline{Y}_N)^۲$، پس $\sigma^۲ = \frac{N-۱}{N}S^۲$، لذا رابطهٔ (۴.۲) به‌صورت زیر درمی‌آید

$$
\mathrm{V}(\overline{Y}_n) = \frac{N-۱}{nN}\cdot\frac{N-n}{N-۱}S^۲ = \frac{N-n}{nN}S^۲ = \left(\frac{۱}{n}-\frac{۱}{N}\right)S^۲
$$

با توجه به $f = \frac{n}{N}$، رابطهٔ بالا را می‌توان به‌صورت زیر هم نوشت

$$
\mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{n}-\frac{۱}{N}\right)S^۲ = \left(۱-\frac{n}{N}\right)\frac{S^۲}{n} = (۱-f)\frac{S^۲}{n} \qquad (۵.۲)
$$

◘

اگر بخواهیم واریانس $\overline{Y}_n$ را از فرمولهای (۴.۲) یا (۵.۲) به‌دست آوریم چون $S^۲$ی جامعه و یا $\sigma^۲$ی جامعه را نداریم، امکان محاسبهٔ آن نیست. خواهیم دید که به کمک نمونه می‌توانیم برآوردی برای واریانس $\overline{Y}_n$ تهیه کنیم. مثال زیر، قضیه بالا را بهتر تشریح می‌کند.

مثال ۴.۲ از جامعهٔ

$$\{۳,\ ۵,\ ۶,\ ۴,\ ۲\}$$

همهٔ نمونه‌های ممکن به‌حجم ۳ را انتخاب کنید و مستقیماً واریانس متغیر تصادفی $\overline{Y}_n$ را به‌دست آورید. سپس درستی قضیهٔ ۴.۲ را با یکی از فرمولهای مربوط بررسی کنید.

در جدول زیر، $\binom{۵}{۳} = ۱۰$ نمونهٔ ممکن و میانگینهای آنها را فهرست کرده‌ایم

نمونه	$\overline{Y}_n$
(۳ ۵ ۶)	$\frac{۱۴}{۳}$
(۳ ۵ ۴)	$\frac{۱۲}{۳}$
(۳ ۵ ۲)	$\frac{۱۰}{۳}$
(۳ ۶ ۴)	$\frac{۱۳}{۳}$
(۳ ۶ ۲)	$\frac{۱۱}{۳}$
(۵ ۶ ۴)	$\frac{۱۵}{۳}$
(۵ ۶ ۲)	$\frac{۱۳}{۳}$
(۶ ۴ ۲)	$\frac{۱۲}{۳}$
(۳ ۴ ۲)	$\frac{۹}{۳}$
(۵ ۴ ۲)	$\frac{۱۱}{۳}$

به‌سهولت می‌توان دید که مطابق قضیهٔ ۲.۲ میانگین $\overline{Y}_n$ برابر میانگین جامعه، یعنی $\overline{Y}_N = ۴$ است.

$$E(\overline{Y}_n) = \frac{۱}{۱۰}\left(\frac{۱۴}{۳} + \cdots + \frac{۱۱}{۳}\right) = \frac{۱}{۱۰}\left(\frac{۱۲۰}{۳}\right) = ۴$$

اگر با توجه به مقادیر $\overline{Y}_n$ و میانگین آن، ۴، مستقیماً واریانس $\overline{Y}_n$ را محاسبه کنیم، نتیجه می‌شود

$$\mathrm{V}(\overline{Y}_n) = \frac{۱}{۱۰}\left[\left(\frac{۱۴}{۳} - ۴\right)^۲ + \cdots + \left(\frac{۱۱}{۳} - ۴\right)^۲\right] = \frac{۱}{۳}$$

اگر بخواهیم درستی رابطهٔ $\mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right) S^۲$ را بررسی کنیم باید ابتدا $S^۲$ی جامعه را حساب کنیم:

$$S^۲ = \frac{۱}{N - ۱}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲ = \frac{۱}{۴}[(۳ - ۴)^۲ + (۵ - ۴)^۲ + \cdots + (۲ - ۴)^۲] = \frac{۵}{۲}$$

پس با توجه به $N = ۵$ و $n = ۳$،

$$\mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{۳} - \frac{۱}{۵}\right)\frac{۵}{۲} = \frac{۱}{۳}$$

همان‌طور که متذکر شدیم، در عمل تنها یکی از نمونه‌های ممکن دردست است و لذا محاسبهٔ $\mathrm{V}(\overline{Y}_n)$ میسر نیست.

۵.۲ تعریف واریانس نمونه

واریانس نمونه‌ای یا واریانس نمونه را که با نماد $s^۲$ نشان می‌دهند با رابطهٔ زیر تعریف می‌کنند

$$s^۲ = \frac{۱}{n-۱}\sum_{i=۱}^{n}(Y_i - \overline{Y}_n)^۲ = \frac{۱}{n-۱}\left[\sum_{i=۱}^{n} Y_i^۲ - n\overline{Y}_n^۲\right]$$

این تعریف، همتای نمونه‌ای تغییرات جامعه، یعنی $S^۲$ است.

قضیهٔ ۵.۲ واریانس نمونهٔ تصادفی بدون جایگذاری به‌حجم n از جامعه‌ای به‌حجم N، برآوردکنندهٔ ناریب $S^۲$ی جامعه است.

برهان. درواقع، اگر واریانس نمونه را با $s^۲$ نشان دهیم، باید ثابت کنیم

$$E(s^۲) = S^۲ \qquad (۶.۲)$$

آمارهٔ $s^۲$ متغیری تصادفی است. بدیهی است تعداد مقادیری که $s^۲$ اختیار می‌کند برابر با تعداد نمونه‌های ممکن، یعنی $\binom{N}{n}$ است. البته ممکن است بعضی از مقادیر تکراری باشند. برای اثبات رابطهٔ بالا، می‌توانیم بنویسیم

$$s^۲ = \frac{۱}{n-۱}\sum_{i=۱}^{n}(Y_i - \overline{Y}_n)^۲ = \frac{۱}{n-۱}\sum_{i=۱}^{n}[(Y_i - \overline{Y}_N) - (\overline{Y}_n - \overline{Y}_N)]^۲$$

میانگین پرانتز اول برابر با مقدار پرانتز دوم است، لذا با استفاده از اتحاد

$$\sum_{i=۱}^{n}(Y_i - \overline{Y}_n)^۲ = \sum_{i=۱}^{۲} Y_i^۲ - n\overline{Y}_n^۲$$

داریم

$$s^۲ = \frac{۱}{n-۱}\left[\sum_{i=۱}^{n}(Y_i - \overline{Y}_N)^۲ - n(\overline{Y}_n - \overline{Y}_N)^۲\right]$$

اگر از دو طرف برابری امید ریاضی بگیریم، نتیجه می‌شود

$$E(s^۲) = \frac{۱}{n-۱}\left[\sum_{i=۱}^{n} E(Y_i - \overline{Y}_N)^۲ - nE(\overline{Y}_n - \overline{Y}_N)^۲\right] \qquad \text{(الف)}$$

اما با استفاده از برابری $(N-۱)S^۲ = N\sigma^۲$

$$E(Y_i - \overline{Y}_N)^۲ = \sigma_Y^۲ = \frac{N-۱}{N}S^۲ \qquad \text{(ب)}$$

و با توجه به اینکه $\overline{Y}_N$ میانگین $\overline{Y}_n$ است

$$E(\overline{Y}_n - \overline{Y}_N)^۲ = \mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right)S^۲ \qquad \text{(ج)}$$

اگر رابطه‌های (ب) و (ج) را در (الف) منظور کنیم، به‌دست می‌آوریم

$$E(s^۲) = \frac{۱}{n-۱}\left[\sum_{i=۱}^{n}\frac{N-۱}{N}S^۲ - n\left(\frac{۱}{n} - \frac{۱}{N}\right)S^۲\right]$$

$$= \frac{۱}{n-۱}\left[\frac{n}{N}(N-۱)S^۲ - \left(۱ - \frac{n}{N}\right)S^۲\right] = S^۲$$

پس $E(s^۲) = S^۲$ و قضیه ثابت می‌شود. □

فرع ۱ . برآوردکنندهٔ نااریب $\mathrm{V}(\overline{Y}_n)$ در یک نمونه‌گیری تصادفی سادهٔ بدون جایگذاری به‌صورت زیر است

$$\hat{\mathrm{V}}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right)s^۲ = (۱-f)\frac{s^۲}{n} \qquad (۷.۲)$$

زیرا طبق قضیهٔ ۵.۲

$$E[\hat{\mathrm{V}}(\overline{Y}_n)] = \left(\frac{۱}{n} - \frac{۱}{N}\right)E(s^۲) = \left(\frac{۱}{n} - \frac{۱}{N}\right)S^۲ = \mathrm{V}(\overline{Y}_n)$$

در رابطهٔ (۷.۲)، $s^۲$ واریانس نمونهٔ به‌حجم n است.

مثال ۵.۲ در مثال قبل دیدیم که اگر همهٔ نمونه‌های ممکن به‌حجم ۳ را از جامعه

$$\{۳,\ ۵,\ ۶,\ ۴,\ ۲\}$$

بگیریم، $V(\overline{Y}_n)$ برابر با $\frac{۱}{۲}$ به‌دست می‌آید. در عمل همهٔ نمونه‌ها را نداریم. فرض کنید تنها نمونه حاصل، نمونهٔ ۴، ۶، ۳ باشد. به‌کمک این نمونه و به‌وسیلهٔ فرمول (۷.۲) می‌توانیم برآوردی ناااریب برای $V(\overline{Y}_n)$ به‌دست آوریم. برای این نمونه $\overline{Y}_n = \frac{۱۳}{۳}$. $s^۲$ی این نمونه را حساب می‌کنیم. چون $n = ۳$

$$s^۲ = \frac{۱}{۲}\left[\left(۳ - \frac{۱۳}{۳}\right)^۲ + \left(۶ - \frac{۱۳}{۳}\right)^۲ + \left(۴ - \frac{۱۳}{۳}\right)^۲\right] = \frac{۷}{۳}$$

پس

$$\hat{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right) s^۲ = \left(\frac{۱}{۳} - \frac{۱}{۵}\right)\frac{۷}{۳} = \frac{۱۴}{۴۵}$$

که به واریانس واقعی $\overline{Y}_n$ که برابر با $\frac{۱}{۲}$ است، نزدیک است. اگر در این مثال برای هر ۱۰ نمونهٔ ممکن $s^۲$ها را حساب کنیم، به‌سهولت می‌بینیم که میانگین ۱۰ مقدار $s^۲$ برابر با $S^۲$ی جامعه است. ▲

فرع ۲ . از برابری (۵.۲) نتیجه می‌شود که

$$\sigma(\overline{Y}_n) = \sqrt{\frac{۱}{n} - \frac{۱}{N}}\, S = \sqrt{\frac{۱ - f}{n}} S \qquad (۸.۲)$$

فرع ۳ . از برابری (۷.۲) نتیجه می‌شود که

$$\hat{\sigma}(\overline{Y}_n) = s\sqrt{\frac{۱}{n} - \frac{۱}{N}} = s\sqrt{\frac{۱ - f}{n}} \qquad (۹.۲)$$

توجه کنید که $\hat{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right) s^۲$ برآوردکنندهٔ ناااریب $V(\overline{Y}_n)$ است، ولی وقتی با گرفتن جذر از طرفین آن به (۹.۲) می‌رسیم $\hat{\sigma}(\overline{Y}_n)$ حاصل، برآوردکنندهٔ ناااریب $\sigma(\overline{Y}_n)$ نیست، زیرا عملگر امید تحت عمل جذر ویژگی ناااریبی را حفظ نمی‌کند. لازم به‌تذکر است که وقتی جامعهٔ مورد نمونه‌گیری دارای توزیع نرمال است، می‌توان برآوردکننده‌ای ناااریب برای $\sigma(\overline{Y}_n)$ یافت.

قضیهٔ ۶.۲ $N\overline{Y}_n$ برآوردکنندهٔ ناااریب T_N، مجموع واحدهای جامعه است.

برهان. بدیهی است که T_N معمولاً مجهول و مقداری ثابت است، و باید به‌کمک نمونه برآوردی برای آن یافت. اگر طرفین برابری

$$E(\overline{Y}_n) = \overline{Y}_N$$

را در N، حجم جامعه، ضرب کنیم، نتیجه می‌شود

$$NE(\overline{Y}_n) = N\overline{Y}_N$$

$N\overline{Y}_N$ برابر T_N است، پس

$$T_N = NE(\overline{Y}_n) = E(N\overline{Y}_n)$$

مفهوم این برابری آن است که $N\overline{Y}_n$ برآوردکنندهٔ ناارِیب T_N است، پس

$$\hat{T}_N = N\overline{Y}_n \qquad (۱۰.۲)$$

و قضیه ثابت می‌شود. □

مثال ۶.۲ از صد قطعه زمین زیر کشت گندم یک روستا، ده قطعه به تصادف و بدون جایگذاری انتخاب کرده‌ایم. میزان محصول این ده قطعه برحسب تن به‌صورت زیرند

$$\{۵ \quad ۱۰ \quad ۱۶ \quad ۲۰ \quad ۱۴ \quad ۱۵ \quad ۲۵ \quad ۵ \quad ۱۸ \quad ۱۲\}$$

مجموع محصولات گندم صد قطعه را برآورد کنید.

از رابطهٔ (۱۰.۲)، یعنی از رابطهٔ $\hat{T}_N = N\overline{Y}_n$ استفاده می‌کنیم

$$\overline{Y}_n = \frac{۱}{۱۰}(۵ + ۱۰ + \cdots + ۱۲) = ۱۴$$

لذا

$$\hat{T}_N = ۱۰۰(۱۴) = ۱۴۰۰$$

پس برآورد ناارِیب مجموع محصولات گندم صد قطعه برابر با ۱۴۰۰ تن است. ▲

فرع ۱ . واریانس برآوردکنندهٔ مجموع واحدهای جامعه عبارت است از

$$\begin{aligned}\sigma^۲(\hat{T}_N) = \mathrm{V}(\hat{T}_N) = \mathrm{V}(N\overline{Y}_n) &= N^۲\mathrm{V}(\overline{Y}_n) = N^۲\left(\frac{۱}{n} - \frac{۱}{N}\right)S^۲ \\ &= (۱ - f)\frac{N^۲S^۲}{n} \qquad (۱۱.۲)\end{aligned}$$

از رابطهٔ بالا، پس از جذرگیری نتیجه می‌شود که

$$\sigma(\hat{T}_N) = NS\sqrt{\frac{۱}{n} - \frac{۱}{N}} = NS \cdot \sqrt{\frac{۱ - f}{n}} \qquad (۱۲.۲)$$

مثال ۷.۲ برای داده‌های مثال ۶.۲، اگر بدانیم که $S^۲$ی جامعه برابر با ۴۱ است، واریانس برآوردکنندهٔ کل محصول گندم صد قطعه را بیابید.

در مثال ۶.۲ برآورد کل محصول گندم برابر با ۱۴۰۰ تن بود. اگر همهٔ نمونه‌های ممکن به‌حجم ۱۰ را از جامعهٔ به‌حجم ۱۰۰ انتخاب کنیم و کل محصول هر نمونه را بیابیم محققاً میانگین محصول همهٔ نمونه‌ها برابر با محصول واقعی گندم صد قطعه می‌شود. در هر حال محصول کل هر نمونه، برآوردی از کل محصول جامعه است. این برآوردها واریانسی دارند که مطابق (۱۱.۲) حساب می‌شود

$$\mathrm{V}(\hat{T}_N) = \left(\frac{۱}{۱۰} - \frac{۱}{۱۰۰}\right)(۱۰۰)^۲(۴۱) = ۳۶۹۰۰$$

$$\sigma(\hat{T}_N \simeq ۱۹۲$$

▲

فرع ۲ . معمولاً $S^۲$ی جامعه در دست نیست و در نتیجه $\mathrm{V}(\hat{T}_N)$ را نمی‌توان محاسبه کرد. لذا باید برآوردی برای آن به‌دست آورد. با توجه به اینکه $s^۲$ی نمونه، برآوردکنندهٔ ناااریب $S^۲$ی جامعه است، داریم

$$\hat{\mathrm{V}}(\hat{T}_N) = \left(\frac{۱}{n} - \frac{۱}{N}\right) N^۲ s^۲ = (۱ - f)\frac{N^۲ s^۲}{n} \qquad (۱۳.۲)$$

این برآوردکننده، ناااریب است، زیرا

$$E[\hat{\mathrm{V}}(\hat{T}_N)] = \left(\frac{۱}{n} - \frac{۱}{N}\right) N^۲ E(s^۲) = \left(\frac{۱}{n} - \frac{۱}{N}\right) N^۲ S^۲ = \mathrm{V}(\hat{T}_N)$$

از (۱۳.۲) نتیجه می‌شود که

$$\hat{\sigma}(\hat{T}_N) = Ns\sqrt{\frac{۱}{n} - \frac{۱}{N}} = Ns\sqrt{\frac{۱ - f}{n}} \qquad (۱۴.۲)$$

بدیهی است $\hat{\sigma}(\hat{T}_N)$، یعنی برآوردکنندهٔ انحراف معیار برای برآوردکنندهٔ مجموع واحدهای جامعه، اریب است، زیرا عمل جذرگیری، حافظ ویژگی ناااریبی نیست.

مثال ۸.۲ برای داده‌های مثال ۶.۲، وقتی $S^۲$ی جامعه مجهول است (که معمولاً چنین است) برآورد واریانس برآوردکنندهٔ کل محصول صد قطعه را بیابید.

برای استفاده از (۱۳.۲)، ابتدا $s^۲$ی نمونه را محاسبه می‌کنیم. قبلاً دیدیم که $\overline{Y}_n = ۱۴$، پس

$$s^۲ = \frac{۱}{۹}[(۱۵ - ۱۴)^۲ + \cdots + (۱۲ - ۱۴)^۲] = ۴۰$$

لذا

$$\hat{\mathrm{V}}(\hat{T}_N) = \left(\frac{۱}{۱۰} - \frac{۱}{۱۰۰}\right)(۱۰۰)^۲(۴۰) = ۳۶۰۰۰$$

پس

$$\hat{\sigma}(\hat{T}_N) \simeq ۱۸۶$$

بدیهی است ۳۶۰۰۰ برآوردی ناا‌ریب برای واریانس، و ۱۸۶ برآوردی اریب برای انحراف معیار است. ▲

۶.۲ ضریب تصحیح واریانس برای جامعهٔ متناهی

در آمار ریاضی دیده‌ایم که وقتی از جامعه‌ای به‌حجم نامتناهی نمونه‌ای تصادفی به‌حجم n می‌گیریم، داریم

$$\mathrm{V}(\overline{Y}_n) = \frac{\sigma^۲}{n} \qquad (۱۵.۲)$$

که در آن، $\sigma^۲$ واریانس جامعه است. وقتی جامعه متناهی و به‌حجم N باشد، بنابر (۴.۲)، داریم

$$\mathrm{V}(\overline{Y}_n) = \frac{\sigma^۲}{n}\frac{N - n}{N - ۱}$$

ملاحظه می‌شود که با منظورکردن ضریب $\frac{N-n}{N-۱}$ در (۱۵.۲)، می‌توان آن را برای جامعهٔ متناهی نیز به‌کار برد. گاهی $N - ۱$ را تقریباً با N برابر می‌گیرند و ضریب $\frac{N-n}{N}$ را در (۱۵.۲) منظور می‌کنند. این ضریب را ضریب تصحیح واریانس جامعهٔ متناهی می‌نامند و آن را با نماد fpc* نشان می‌دهند. اگر کسر نمونه‌گیری، $f = \frac{n}{N}$، کوچک باشد، غالباً همان $\frac{N-n}{N-۱}$ را به‌کار می‌برند. وقتی N بزرگ است و n بزرگ نیست، این ضریب نزدیک ۱ است و بر انحراف معیار و واریانس میانگین نمونه اثری چندان ندارد. در عمل وقتی f از ۰٫۰۵ تجاوز نمی‌کند ضریب تصحیح را نادیده می‌گیرند. نادیده گرفتن ضریب تصحیح واریانس، معمولاً موجب بیش برآورد انحراف معیار $\overline{Y}_n$ می‌شود.

* fpc نمادی برای اصطلاح finite population correction است.

۷.۲ حدود اطمینان میانگین جامعه

قبل از بحث در حدود اطمینان میانگین جامعه، متذکر می‌شویم که در نمونه‌گیری تصادفی ساده بدون جایگذاری، تعداد $\overline{Y}_n$های حاصل از نمونه‌های ممکن، وقتی حجم جامعه نسبتاً بزرگ است بسیار زیاد است. مثلاً اگر از جامعهٔ به‌حجم ۵۰ نمونه‌ای تصادفی به‌حجم ۱۰ بگیریم، تعداد $\overline{Y}_n$ها برابر با $\binom{۵۰}{۱۰} = ۱۰۲۷۲۲۷۸۱۷۰$ است. می‌بینید که با وجود کوچک بودن حجم جامعه و نمونه، تعداد مقادیری که متغیر تصادفی $\overline{Y}_n$ اختیار می‌کند از ۱۰ میلیارد متجاوز است. اگر فراوانیهای مقادیر $\overline{Y}_n$ را مشخص کنیم و اگر دامنهٔ تغییر مقادیر $\overline{Y}_n$ را به بازه‌هایی تقسیم کرده و فراوانی هر بازه را و سپس بافتنمای $\overline{Y}_n$ را به‌دست آوریم تقریباً به نموداری شبه نرمال می‌رسیم. ثابت شده است که در هر جامعه با انحراف معیار متناهی، توزیع میانگین نمونه، وقتی حجم نمونه بزرگ می‌شود، به توزیع نرمال می‌گراید. این ویژگی برای جامعهٔ با حجم نامتناهی وجود دارد ولی برای جامعه‌های متناهی نیز، تحت شرایطی که غالباً برقرارند، به‌صورت تقریبی موجود است. بر این مبناست که معمولاً فرض می‌کنند $\overline{Y}_n$ و در نتیجه T_n حول مقدار همتای جامعه‌ای آنها، یعنی $\overline{Y}_N$ و T_N به‌صورت نرمال توزیع می‌شوند. اگر این فرض برقرار باشد به‌کمک توزیع نرمال می‌توان حدود اطمینان میانگین جامعه و مجموع واحدهای جامعه را از روی توزیع نمونه‌ای میانگین به‌دست آورد.
چون $\overline{Y}_n \sim N(\overline{Y}_N,\ V(\overline{Y}_n))$، نتیجه می‌شود که $Z = \frac{\overline{Y}_n - \overline{Y}_N}{\sigma(\overline{Y}_n)} \sim N(۰,\ ۱)$. اگر قرار دهیم $\underset{Z}{\phi}(z) = ۱ - \frac{\alpha}{۲}$، که در آن ϕ تابع توزیع نرمال استاندارد است، آن‌گاه

$$۱ - \alpha = \Pr(-z < Z < z) = \Pr\left(-z < \frac{\overline{Y}_n - \overline{Y}_N}{\sigma(\overline{Y}_n)} < z\right)$$

با توجه به $\sigma(\overline{Y}_n) = S\sqrt{\frac{۱-f}{n}}$ از رابطهٔ بالا به‌سهولت حدود اطمینان پایین و بالای $\overline{Y}_N$ با ضریب اطمینان $۱ - \alpha$، به‌صورت زیر به‌دست می‌آیند

$$\begin{cases} \overline{Y}_N \text{ حد پایین اطمینان} = \overline{Y}_L = \overline{Y}_n - zS\sqrt{\dfrac{۱-f}{n}} \\ \overline{Y}_N \text{ حد بالای اطمینان} = \overline{Y}_U = \overline{Y}_n + zS\sqrt{\dfrac{۱-f}{n}} \end{cases} \qquad (۱۶.۲)$$

و با ضرب دو طرف این برابریها در N، حدود اطمینان پایین و بالای T_N با ضریب اطمینان $۱ - \alpha$ عبارت‌اند از

$$\begin{cases} T_N \text{ حد پایین اطمینان} = T_L = N\overline{Y}_n - zSN\sqrt{\dfrac{۱-f}{n}} \\ T_N \text{ حد بالای اطمینان} = T_U = N\overline{Y}_n + zSN\sqrt{\dfrac{۱-f}{n}} \end{cases} \qquad (۱۷.۲)$$

لازم به‌ذکر است که وقتی حجم نمونه از ۵۰ بیشتر باشد، $\overline{Y}_n$، متغیر تصادفی نرمال با میانگین $\overline{Y}_N$ و انحراف معیار $S\sqrt{\frac{1-f}{n}}$ است. شبیه‌سازیها نشان می‌دهند که حتی برای $n = ۲۰$ هم تقریب خیلی بد نیست.

با توجه به جدول توزیع نرمال، مقدار z در حدود مذکور، به‌ازای برخی احتمالهای متداول عبارت‌اند از

احتمال (٪)	۵۰	۸۰	۹۰	۹۵	۹۹
مقدار z	۰٫۶۷	۱٫۲۸	۱٫۶۴	۱٫۹۶	۲٫۵۸

اگر حجم نمونه از ۵۰ کمتر باشد مقادیر z را به‌جای استفاده از توزیع نرمال، از جدول توزیع استودنت با $(n - ۱)$ درجهٔ آزادی به‌دست می‌آورند. $n - ۱$، درجهٔ آزادی واریانس برآوردشدهٔ $s^۲$ است. توزیع استودنت وقتی دقیقاً صادق است که مشاهدات Y_i دارای توزیع نرمال بوده و N نامتناهی باشد. البته کمی انحراف از نرمال بودن اثری چندان بر شیوهٔ کار ندارد. در نمونه‌های کوچک با توزیعهای خیلی چاوله، برای محاسبهٔ حدود اطمینان، روشهایی خاص ضروری‌اند.

توجه در محاسبهٔ حدود اطمینان، معمولاً $S^۲$ مجهول است، لذا معمولاً از برآورد نمونه‌ای آن، $s^۲$ استفاده می‌کنند. در نتیجه به‌جای حدود اطمینان، برآوردهای حدود اطمینان به‌دست می‌آیند.

مثال ۹.۲ دانشکده‌ای ۴۸۴ دانشجو دارد. برای تعیین متوسط مخارج ماهیانهٔ آنها، نمونه‌ای به حجم $n = ۹$ با روش تصادفی بدون جایگذاری انتخاب می‌کنیم. مخارج ماهیانهٔ آنها به‌شرح زیر است:

$$Y_۱ = ۳۳۵۰ \qquad Y_۴ = ۴۳۰۰ \qquad Y_۷ = ۴۵۰۰$$

$$Y_۲ = ۳۲۰۰ \qquad Y_۵ = ۴۰۰۰ \qquad Y_۸ = ۴۲۵۰$$

$$Y_۳ = ۵۲۰۰ \qquad Y_۶ = ۴۱۰۰ \qquad Y_۹ = ۳۹۰۰$$

برآورد میانگین مخارج ماهیانهٔ دانشجویان دانشکده را بیابید و بازهٔ اطمینانی برای میانگین مخارج ماهیانه، با ضریب اطمینان ۹۵٪ به‌دست آورید.

به‌سهولت، از روی نمونه، به‌دست می‌آوریم

$$\sum_{i=۱}^{۹} Y_i = ۳۶۸۰۰ \quad , \quad \sum_{i=۱}^{۹} Y_i^۲ = ۱۵۳۳۲۵۰۰۰$$

پس $\overline{Y}_n = \frac{۳۶۸۰۰}{۹} \simeq ۴۰۸۹$ لذا $\hat{\overline{Y}}_N = \overline{Y}_n \simeq ۴۰۸۹$ که برآورد ناریب $\overline{Y}_N$، یعنی متوسط

مخارج ماهیانه هر دانشجوست. برای تعیین بازهٔ اطمینان، ابتدا $s^۲$ی نمونه را حساب می‌کنیم

$$\begin{aligned} s^۲ &= \frac{۱}{n-۱}\sum_{i=۱}^{n}(Y_i - \overline{Y}_n)^۲ = \frac{۱}{n-۱}(\sum_{i=۱}^{n} Y_i^۲ - n\overline{Y}_n^۲) \\ &= \frac{۱}{n-۱}\left[\sum_{i=۱}^{n} Y_i^۲ - \frac{(\sum_{i=۱}^{n} Y_i)^۲}{n}\right] \\ &= \frac{۱}{۸}\left[۱۵۳۳۲۵۰۰۰ - \frac{(۳۶۸۰۰)^۲}{۹}\right] \simeq ۳۵۶۷ \end{aligned}$$

پس

$$s \simeq ۵۹{,}۷۲$$

با قبول فرض نرمال بودن توزیع $\overline{Y}_n$، از (۱۶.۲) استفاده می‌کنیم و به‌جای S، برآورد (اریب) آن s را قرار می‌دهیم، در نتیجه حدود اطمینان $\overline{Y}_N$، با توجه به $z_{۰{,}۹۷۵} = ۱{,}۹۶$ به‌صورت زیر به‌دست می‌آیند:

$$\begin{cases} \hat{\overline{Y}}_L = \overline{Y}_n - zs\sqrt{\dfrac{۱-f}{n}} = ۴۰۸۹ - \dfrac{۱{,}۹۶}{۳}(۵۹{,}۷۲)\sqrt{۱-\dfrac{۹}{۴۸۴}} \simeq ۴۰۵۰{,}۷۰۹ \\ \hat{\overline{Y}}_U = \overline{Y}_n + zs\sqrt{\dfrac{۱-f}{n}} = ۴۰۸۹ + \dfrac{۱{,}۹۶}{۳}(۵۹{,}۷۲)\sqrt{۱-\dfrac{۹}{۴۸۴}} \simeq ۴۱۲۷{,}۲۹۱ \end{cases}$$

یعنی $\overline{Y}_N$، متوسط واقعی مخارج ماهیانه، با احتمال ۰٫۹۵ در بازهٔ (۴۰۵۰٫۷۰۹, ۴۱۲۷٫۲۹۱) قرار دارد. ▲

مثال ۱۰.۲ براساس داده‌های مثال ۶.۲، بازهٔ اطمینانی برای کل محصول گندم صد قطعه زمین زیر کشت، با ضریب اطمینان ۰٫۹۵ به‌دست آورید.

در مثال ۶.۲ داشتیم $n = ۱۰$، $\overline{Y}_n = ۱۴$ و براساس محاسبات مثال ۸.۲ داشتیم $s^۲ = ۴۰$، پس $s \simeq ۶{,}۳$. با توجه به (۱۷.۲) و قرار دادن s به‌جای S، برآورد حدود اطمینان، وقتی $۱-\alpha = ۰{,}۹۵$، عبارت‌اند از

$$\hat{T}_L = N\overline{Y}_n - zNs\sqrt{\frac{۱-f}{n}} = ۱۰۰(۱۴) - (۱{,}۹۶)(۱۰۰)(۶{,}۳)\sqrt{\frac{۹}{۱۰۰}} \simeq ۱۰۲۹{,}۵۶$$

$$\hat{T}_U = N\overline{Y}_n + zNs\sqrt{\frac{۱-f}{n}} = ۱۰۰(۱۴) + (۱{,}۹۶)(۱۰۰)(۶{,}۳)\sqrt{\frac{۹}{۱۰۰}} \simeq ۱۷۷۰{,}۴۴$$

یعنی محصول کل صد قطعه زمین با احتمال ۰٫۹۵ در بازهٔ (۱۰۲۹٫۵۶, ۱۷۷۰٫۴۴) قرار دارد. ▲

قبل از اینکه به نمونه‌گیری تصادفی ساده با جایگذاری بپردازیم دو قضیه در خصوص نمونه‌گیری تصادفی سادهٔ بدون جایگذاری که در کاربردها مفیدند مطرح می‌کنیم.

قضیهٔ ۷.۲ زیر نمونهٔ تصادفی سادهٔ بدون جایگذاری به‌حجم n از نمونهٔ تصادفی سادهٔ بدون جایگذاری به‌حجم $n^*(n^* > n)$، خود نمونه‌ای تصادفی بدون جایگذاری به حجم n از جامعهٔ اصلی است.

برهان. اگر مجموعهٔ اعضای نمونه تصادفی به‌حجم n^* را با مجموعهٔ S^* نشان دهیم، احتمال وقوع هر S^* معین عبارت است از

$$\Pr(S^*) = \frac{۱}{\binom{N}{n^*}}$$

که در آن، N حجم جامعه است. اگر نمونهٔ به‌حجم n را که از S^* به‌تصادفی و بدون جایگذاری انتخاب می‌کنیم با $\{Y_۱,\ Y_۲, \ldots , Y_n\}$ نشان دهیم، آن‌گاه بنابر فرمول

$$\Pr(\{Y_۱,\ Y_۲, \ldots , Y_n\}) = \sum \Pr(S^*)\Pr(\{Y_۱,\ Y_۲, \ldots , Y_n\}|S^*)$$

که مجموعیابی روی همهٔ آن S^*هایی است که شامل $\{Y_۱,\ Y_۲, \ldots , Y_n\}$ هستند. بدیهی است که $\Pr(\{Y_۱,\ Y_۲, \ldots , Y_n\}) = ۰$ اگر $\{Y_۱,\ Y_۲, \ldots , Y_n\}$ زیرمجموعهٔ S^* نباشد و اگر $\{Y_۱,\ Y_۲, \ldots , Y_n\}$ زیرمجموعهٔ S^* باشد $\Pr(\{Y_۱,\ Y_۲, \ldots , Y_n\}) = \frac{۱}{\binom{n^*}{n}}$. لذا بنابر فرمول بالا

$$\Pr(\{Y_۱, Y_۲, \ldots , Y_n\}) = \frac{۱}{\binom{N}{n^*}} \cdot \frac{۱}{\binom{n^*}{n}} \cdot (\text{تعداد } S^* \text{هایی که شامل } \{Y_۱, Y_۲, \ldots , Y_n\} \text{اند})$$

اما برای تعیین S^*هایی که شامل اعضای معین و ثابت $(Y_۱,\ Y_۲, \ldots , Y_n)$ هستند باید توجه داشت که تنها $n^* - n$ عضو آنها از بین $N - n$ عضو باقیمانده جامعه به تصادف و بدون جایگذاری انتخاب می‌شود که تعداد آنها برابر است با

$$\binom{N-n}{n^*-n}$$

پس

$$\Pr(\{Y_۱,\ Y_۲, \ldots , Y_n\}) = \frac{۱}{\binom{N}{n^*}} \cdot \frac{۱}{\binom{n^*}{n}} \cdot \binom{N-n}{n^*-n} = \frac{۱}{\binom{N}{n}}$$

یعنی، نمونهٔ $\{Y_۱,\ Y_۲, \ldots , Y_n\}$ نمونه‌ای تصادفی به‌حجم n و بدون جایگذاری از جامعهٔ اصلی است. □

قضیهٔ ۸.۲ ترکیب دو نمونهٔ تصادفی ساده، اولی به‌حجم $n_۱$ از جامعهٔ اصلی، و دومی به‌حجم $n_۲$ از زیرجامعهٔ $N - n_۱$ که شامل واحدهای نمونهٔ اول نیست، نمونهٔ تصادفی ساده‌ای به‌حجم $n_۱ + n_۲$ از جامعهٔ اصلی است.

برهان قضیه شبیه برهان قضیهٔ ۷.۲ است که آن را به‌عهدهٔ خواننده می‌گذاریم. □

۸.۲ نمونه‌گیری تصادفی سادهٔ با جایگذاری

اگر در انتخاب n واحد نمونه از جامعهٔ به‌حجم N، پس از انتخاب هر واحد، و ثبت مشخصهٔ آن، آن را به جامعه برگردانیم و سپس انتخاب بعدی را انجام دهیم نمونه‌گیری تصادفی را با جایگذاری می‌نامند. در این روش، انتخاب هر واحد از انتخاب واحدهای دیگر مستقل است، و احتمال استخراج هر واحد مشخص، برابر با $\frac{۱}{N}$ است، زیرا حجم جامعه در هر انتخاب، ثابت و برابر N باقی می‌ماند. هر واحد نمونه به N راه انتخاب می‌شود، یعنی هر واحد نمونه ممکن است $Y_۱$ یا $Y_۲$، . . . ، یا Y_N باشد. پس برای انتخاب یک نمونه به‌حجم n، $\underbrace{N \cdot N \cdots\cdots N}_{n \text{ بار}} = N^n$ راه وجود دارد. بنابراین تعداد نمونه‌های ممکن تصادفی به‌روش با جایگذاری و با حجم n برابر با N^n است. از طرفی،

$$\begin{aligned}&\Pr(\text{انتخاب یک نمونهٔ مشخص به‌حجم } n)\\ &= \Pr(\text{انتخاب واحد اول}) \cdot \Pr(\text{انتخاب واحد دوم}) \cdots \Pr(\text{انتخاب واحد } n\text{ام}) =\\ &= \frac{۱}{N} \cdot \frac{۱}{N} \cdots \frac{۱}{N} = \frac{۱}{N^n}\end{aligned}$$

بدین ترتیب، احتمال رخداد هر یک از N^n نمونهٔ تصادفی با جایگذاری $\frac{۱}{N^n}$ است. در صورتی که احتمال رخداد هر نمونهٔ تصادفی به‌حجم n با روش بدون جایگذاری $۱/\binom{N}{n}$ است. در نمونه‌گیری با جایگذاری احتمال تکراری بودن مشاهدات صفر نیست و هر واحد ممکن است i بار، $۱ \leq i \leq n$، در نمونه ظاهر شود.

مثال ۱۱.۲ در جامعهٔ به‌حجم ۳ همهٔ نمونه‌های تصادفی سادهٔ با جایگذاری و به‌حجم ۲ را مشخص کنید.

اگر واحدهای جامعه را با $u = \{۱, ۲, ۳\}$ نشان دهیم، تعداد نمونه‌های ممکن به روش با جایگذاری $۹ = ۳^۲$ است. نمونه‌های ممکن عبارت‌اند از

$$\begin{array}{ccc}(۱, ۱) & (۲, ۱) & (۳, ۱)\\ (۱, ۲) & (۲, ۲) & (۳, ۲)\\ (۱, ۳) & (۲, ۳) & (۳, ۳)\end{array}$$

▲

تبصره. در نمونه‌گیری تصادفی سادهٔ با جایگذاری، در بعضی از نمونه‌های به‌حجم n همهٔ واحدها متفاوت‌اند، ولی در یک نمونه به‌حجم n امکان دارد فقط k واحد $(k = ۱, ۲, \ldots, n)$ متفاوت باشند. به‌وسیلهٔ استفاده از آنالیز ترکیبیاتی و اعداد استرلینگ ثابت می‌شود احتمال اینکه در یک نمونه به‌حجم n از جامعه به‌حجم N، دارای k واحد متفاوت باشیم برابر است با

$$\Pr(k) = \frac{N^{(k)} S(n,\ k)}{N^n}$$
$$k = ۱,\ ۲, \ldots, n$$

که در آن $S(n,\ k)$ عدد نوع دوم استرلینگ است که با رابطهٔ زیر تعریف می‌شود

$$x^n = \sum_{k=۰}^{n} S(n,\ k) x^{(k)}$$
$$x^{(k)} = x(x-۱)\cdots(x-k+۱)$$
$$N^{(k)} = N(N-۱)\cdots(N-k+۱)$$

۹.۲ میانگین نمونهٔ تصادفی سادهٔ با جایگذاری

فرض می‌کنیم فراوانی واحد iام جامعه در نمونه Z_i و مقدار متغیر برای آن Y_i باشد، یعنی واحد Y_iی جامعه Z_i بار در نمونه ظاهر شود. بدیهی است Z_i، $۱ \le i \le N$ می‌تواند مقادیر ۰ تا n را اختیار کند. با این فرض میانگین نمونه برابر است با

$$\overline{Y}_n = \frac{۱}{n}\sum_{i=۱}^{N} Z_i Y_i \quad \text{و} \quad \sum_{i=۱}^{N} Z_i = n \qquad (۱۸.۲)$$

قضیهٔ ۹.۲ میانگین نمونهٔ تصادفی سادهٔ با جایگذاری، برآوردکنندهٔ نااریب میانگین جامعه است.

برهان. از روابط (۱۸.۲)، داریم

$$E(\overline{Y}_n) = \frac{۱}{n} E\left[\sum_{i=۱}^{N} Z_i Y_i\right] = \frac{۱}{n}\sum_{i=۱}^{N} E(Z_i Y_i)$$

در طرف راست آخرین برابری، Y_iها، همهٔ مقادیر واحدهای جامعه‌اند ولی Z_i متغیری تصادفی است که مقادیر ۰ تا n را اختیار و از توزیع دوجمله‌ای پیروی می‌کند. در این توزیع تعداد امتحانهای برنولی برابر n و احتمال پیروزی در هر امتحان $\frac{۱}{N}$ است. لذا $E(Z_i) = \frac{n}{N}$. پس برابری بالا به‌صورت زیر درمی‌آید

$$E(\overline{Y}_n) = \frac{۱}{n}\sum_{i=۱}^{N} Y_i E(Z_i) = \frac{۱}{n}\sum_{i=۱}^{N} Y_i \cdot \frac{n}{N} = \overline{Y}_N$$

یعنی $\overline{Y}_n$، مثل حالت نمونهٔ تصادفی بدون جایگذاری، برآوردکنندهٔ نااریب میانگین جامعه است. □

مثال ۱۲.۲ از جامعهٔ به‌حجم ۳ به‌صورت $Y_1 = 2$، $Y_2 = 3$، $Y_3 = 4$ همهٔ نمونه‌های ممکن با جایگذاری به‌حجم ۲ را مشخص کنید و درستی برابری $E(\overline{Y}_n) = \overline{Y}_N$ را تحقیق نمایید.

تعداد نمونه‌های ممکن $9 = 3^2$ است. همهٔ نمونه‌ها و میانگین آنها به‌صورت زیر فهرست می‌شوند

میانگین	نمونه	میانگین	نمونه	میانگین	نمونه
۳	۲ ۴	$\frac{5}{2}$	۲ ۳	۲	۲ ۲
$\frac{7}{2}$	۳ ۴	۳	۳ ۳	$\frac{5}{2}$	۳ ۲
۴	۴ ۴	$\frac{7}{2}$	۴ ۳	۳	۴ ۲

بنابراین، با توجه به ۹ مقدار $\overline{Y}_n$، داریم

$$E(\overline{Y}_n) = \frac{1}{9}\left[2 + \frac{5}{2} + \cdots + 4\right] = 3$$

از سوی دیگر

$$\overline{Y}_N = \frac{1}{3}(2 + 3 + 4) = 3$$

لذا برابری مورد نظر برقرار است. ▲

۱۰.۲ واریانس میانگین نمونهٔ تصادفی سادهٔ با جایگذاری

در نمونه‌گیری تصادفی سادهٔ با جایگذاری، N^n میانگین حاصل (که ممکن است بعضی از آنها تکراری باشند) حول $\overline{Y}_N$ پراکنده‌اند. اگر واریانس $\overline{Y}_n$ کم باشد، حتی مقداری از $\overline{Y}_n$ که نسبت به $\overline{Y}_N$ وضعیتی فرین دارد به $\overline{Y}_N$ نزدیک خواهد بود و لذا اگر در عمل تنها با آن مواجه شویم برآورد نااریب نسبتاً مطلوبی خواهد بود. ولی اگر واریانس $\overline{Y}_n$؛ یعنی میزان پراکندگی $\overline{Y}_n$ها زیاد باشد ممکن است در عمل با مقداری از $\overline{Y}_n$ مواجه شویم که به فاصلهٔ زیاد از $\overline{Y}_N$ قرار دارد. در این صورت $\overline{Y}_n$ حاصل از نمونه ضمن اینکه نااریب است، نادقیق است. لذا محاسبهٔ واریانس $\overline{Y}_n$ برای ارزیابی دقت برآورد ضروری است. اگر Z_i مثل بالا، معرف تعداد دفعاتی باشد که واحد iام جامعه در نمونهٔ با جایگذاری ظاهر می‌شود، چون Z_i از توزیع دوجمله‌ای با احتمال پیروزی $\frac{1}{N}$ پیروی می‌کند، داریم

$$\mathrm{V}(Z_i) = n \cdot \frac{1}{N}\left(1 - \frac{1}{N}\right) = \frac{n}{N}\frac{N-1}{N} \qquad \text{(الف)}$$

از طرفی بردار متغیرهای Z_i شامل $Z_۱$ تا Z_N از توزیع چندجمله‌ای پیروی می‌کند (چرا؟) و در این توزیع

$$\operatorname{cov}(Z_i,\ Z_j) = \frac{-n}{N^۲} \qquad (ب)$$

با این مقدمه، می‌توانیم واریانس $\overline{Y}_n$ را حساب کنیم

$$\begin{aligned}
\mathrm{V}(\overline{Y}_n) &= \mathrm{V}\left(\frac{۱}{n}\sum_{i=۱}^{N} Z_i Y_i\right) = \frac{۱}{n^۲}\mathrm{V}(Y_۱ Z_۱ + Y_۲ Z_۲ + \cdots + Y_N Z_N) \\
&= \frac{۱}{n^۲}\left[\sum_{i=۱}^{N}\mathrm{V}(Y_i Z_i) + ۲\sum_{\substack{i,j \\ i<j}}\operatorname{cov}(Y_i Z_i,\ Y_j Z_j)\right] \\
&= \frac{۱}{n^۲}\left[\sum_{i=۱}^{N} Y_i^۲\mathrm{V}(Z_i) + ۲\sum_{\substack{i,j \\ i<j}}\operatorname{cov}(Y_i Z_i,\ Y_j Z_j)\right]
\end{aligned}$$

با توجه به برابریهای (الف) و (ب)، داریم

$$\begin{aligned}
\mathrm{V}(\overline{Y}_n) &= \frac{۱}{n^۲}\left[\sum_{i=۱}^{N} Y_i^۲ \frac{n}{N}\frac{N-۱}{N} + ۲\sum_{\substack{i,j \\ i<j}}\frac{-n}{N^۲}Y_i Y_j\right] \\
&= \frac{N-۱}{nN^۲}\sum_{i=۱}^{N} Y_i^۲ - \frac{۲}{nN^۲}\sum_{\substack{i,j \\ i<j}} Y_i Y_j \\
&= \frac{۱}{nN}\left[\frac{N-۱}{N}\sum_{i=۱}^{N} Y_i^۲ - \frac{۲}{N}\sum_{\substack{i,j \\ i<j}} Y_i Y_j\right]
\end{aligned}$$

از طرفی می‌دانیم

$$(Y_۱ + Y_۲ + \cdots + Y_N)^۲ = \sum_{i=۱}^{N} Y_i^۲ + ۲\sum_{\substack{i,j \\ i<j}} Y_i Y_j$$

اگر از این برابری، مقدار $\sum\limits_{\substack{i,j\\ i<j}} Y_iY_j$ را به‌دست آورده و در رابطهٔ $V(\overline{Y}_n)$ قرار دهیم نتیجه می‌شود

$$\begin{aligned}
V(\overline{Y}_n) &= \frac{1}{nN}\left[\frac{N-1}{N}\sum_{i=1}^{N} Y_i^2 + \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{N}(Y_1+\cdots+Y_N)^2\right]\\
&= \frac{1}{nN}\left[\sum_{i=1}^{N} Y_i^2 - \frac{1}{N}(Y_1+\cdots+Y_n)^2\right]\\
&= \frac{1}{nN}\left[\sum_{i=1}^{N} Y_i^2 - \frac{1}{N}(N\overline{Y}_N)^2\right]\\
&= \frac{1}{nN}\left[\sum_{i=1}^{N} Y_i^2 - N\overline{Y}_N^2\right] = \frac{1}{nN}\sum_{i=1}^{N}(Y_i-\overline{Y}_N)^2 = \frac{\sigma^2}{n}
\end{aligned}$$

که در آن، σ^2 واریانس جامعه است، لذا در نمونه‌گیری تصادفی سادهٔ با جایگذاری

$$V(\overline{Y}_n) = \frac{\sigma^2}{n} \quad , \quad \sigma(\overline{Y}_n) = \frac{\sigma}{\sqrt{n}} \qquad (۱۹.۲)$$

به‌طوری که می‌دانید برای جامعه‌های نامتناهی، وقتی نمونه‌ای به‌حجم n به تصادف انتخاب می‌شود همین ویژگی وجود دارد.

قضیهٔ ۱۰.۲ در نمونه‌گیری تصادفی سادهٔ با جایگذاری، s^2ی نمونه، برآوردکنندهٔ نااریب واریانس جامعه است، یعنی

$$E(s^2) = \sigma^2$$

برهان. می‌دانیم

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(Y_i-\overline{Y}_N)-(\overline{Y}_n-\overline{Y}_N)]^2$$

چون $\overline{Y}_n-\overline{Y}_N$ سیانگین $Y_i-\overline{Y}_N$ است، پس

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i-\overline{Y}_N)^2 - n(\overline{Y}_n-\overline{Y}_N)^2\right]$$

بنابراین، داریم

$$E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}E((Y_i-\overline{Y}_N)^2) - nE((\overline{Y}_n-\overline{Y}_N)^2)\right]$$

اما

$$E((\overline{Y}_n - \overline{Y}_N)^۲) = \mathrm{V}(\overline{Y}_n) \quad , \quad E((Y_i - \overline{Y}_N)^۲) = \sigma^۲$$

لذا

$$E(s^۲) = \frac{۱}{n-۱}\left[\sum_{i=۱}^{n} \sigma^۲ - n\mathrm{V}(\overline{Y}_n)\right]$$

اما، در نمونه‌گیری با جایگذاری

$$\mathrm{V}(\overline{Y}_n) = \frac{\sigma^۲}{n}$$

بنابراین

$$E(s^۲) = \frac{۱}{n-۱}\left[n\sigma^۲ - n\frac{\sigma^۲}{n}\right] = \frac{\sigma^۲}{n-۱}(n-۱)$$
$$= \sigma^۲$$

و قضیه ثابت می‌شود. □

فرع. چون $s^۲$ برآوردکنندهٔ نااریب $\sigma^۲$ است، پس

$$\hat{\mathrm{V}}(\overline{Y}_n) = \frac{s^۲}{n} \qquad (۲۰.۲)$$

این برآوردکننده نااریب است، زیرا

$$E[\hat{\mathrm{V}}(\overline{Y}_n)] = E\left(\frac{s^۲}{n}\right) = \frac{\sigma^۲}{n} = \mathrm{V}[\overline{Y}(n)]$$

در عمل، محاسبهٔ واریانس $\overline{Y}_n$ ممکن نیست، زیرا $\sigma^۲$ مجهول است، ولی از روی نمونه و به کمک $s^۲$ی نمونه می‌توان به‌وسیلهٔ (۲۰.۲) برآورد نااریب $\mathrm{V}(\overline{Y}_n)$ را به‌دست آورد.

تذکر. می‌توان نشان داد که نمونه‌گیری تصادفی سادهٔ با جایگذاری، دقّتی کمتر از نمونه‌گیری تصادفی سادهٔ بدون جایگذاری دارد، زیرا اگر حجم هر دو نمونه را n فرض کنیم واریانس $\overline{Y}_n$ در دو حالت به‌صورت زیرند

$$\text{بدون جایگذاری} : \mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N}\right) S^۲ = \frac{N-n}{nN} S^۲$$

$$\text{با جایگذاری} : \mathrm{V}(\overline{Y}_n) = \frac{\sigma^۲}{n} = \frac{N-۱}{nN} S^۲$$

چون برای $n > ۱$ همواره $N - n < N - ۱$، پس برای $n \neq ۱$ همیشه واریانس برآوردکننده‌ای از میانگین جامعه که با روش تصادفی ساده بدون جایگذاری به‌دست می‌آید کوچکتر از واریانس برآوردکنندهٔ میانگین جامعه به‌روش تصادفی سادهٔ با جایگذاری است. به این دلیل است که معمولاً از روش اول استفاده می‌کنند. لازم به تذکر است که از نمونه‌گیری با جایگذاری در روشی که به‌نام خودگردان (بوت استراپ)[1] موسوم است و برای برآورد تجربی دقیقتر پارامترها به‌کار می‌رود استفاده می‌شود.

مثال ۱۳.۲ نمونه‌ای تصادفی به‌حجم ۵ و به روش با جایگذاری از جامعه‌ای به‌حجم ۲۰ انتخاب کرده‌ایم. اگر مقادیر متغیر تحت بررسی در نمونه به‌صورت زیر باشد

$$(۹ \quad ۴ \quad ۲ \quad ۳ \quad ۲)$$

برآورد میانگین جامعه و برآورد واریانس برآوردکنندهٔ میانگین جامعه را حساب کنید. در این مثال، حجم جامعه ۲۰ و حجم نمونه ۵ است، پس

$$\overline{Y}_n = \hat{\overline{Y}}_N = \frac{۱}{۵}(۲ + ۳ + ۲ + ۴ + ۹) = ۴$$

برای محاسبهٔ برآورد واریانس میانگین نمونه، $\overline{Y}_n$، که مقدار ۴ رخدادی از آن است، ابتدا $s^۲$ی نمونه را حساب می‌کنیم

$$s^۲ = \frac{۱}{۵ - ۱}[(۲ - ۴)^۲ + \cdots + (۹ - ۴)^۲] = ۸٫۵$$

پس

$$\hat{\mathrm{V}}(\overline{Y}_n) = \frac{s^۲}{n} = \frac{۸٫۵}{۵} = ۱٫۷$$

بدیهی است که این برآورد، نااریب است. اگر برآورد انحراف معیار $\overline{Y}_n$ را بخواهیم

$$\hat{\sigma}(\overline{Y}_n) = \sqrt{۱٫۷} \# ۱٫۳$$

این برآورد، اریب است، زیرا عملگر امید ریاضی نسبت به عمل جذر پایا نیست. ▲

۱۱.۲ نسبتها و درصدها

گاهی اوقات بعضی واحدهای جامعه واجد صفت معینی هستند و مایلیم نسبت یا درصد این واحدها را در کل جامعه برآورد کنیم. گاهی نیز دربارهٔ واحدهای جامعه رده‌بندی خاص تعریف

1. Bootstrap

می‌شود به‌صورتی که هر واحد جامعه تنها در یکی از این رده‌ها قرار می‌گیرد و مایلیم نسبت یا درصد واحدهایی از جامعه را که در هر رده قرار می‌گیرند برآورد کنیم. مثلاً اطلاع از نسبت افراد بیکار، یا تعیین درصد بیکاران در جامعهٔ افراد ۲۰ تا ۳۰ ساله یک شهر برای برنامه‌ریزیهای اقتصادی ضروری است.

ابتدا فرض می‌کنیم هر واحد جامعه با توجه به صفتی خاص در یکی از دو رده قرار گیرد. هدف این است که از جامعه، نمونه‌ای به‌روش تصادفی ساده بدون جایگذاری استخراج و به‌کمک آن برآورد کنیم که چه نسبتی و یا چه درصدی از واحدهای جامعه در این دو رده قرار دارند. در این بخش نحوهٔ محاسبهٔ این برآوردها را مطرح می‌کنیم.

۱.۱۱.۲ نمادگذاری

دو رده را با C و $\overline{C}$ نشان می‌دهیم که مکمل هم هستند. هر واحد جامعه در تنها یکی از دو ردهٔ C یا $\overline{C}$ قرار می‌گیرد. حجم جامعه و نمونه را به‌ترتیب، مثل معمول، با N و n نشان می‌دهیم نمادهای زیر را در نظر می‌گیریم

$A =$ تعداد واحدهایی از جامعه که در ردهٔ C هستند

$a =$ تعداد واحدهایی از نمونه که در ردهٔ C هستند

$P = \dfrac{A}{N}$ نسبت تعداد واحدهایی از جامعه که در ردهٔ C هستند

$p = \dfrac{a}{n}$ نسبت تعداد واحدهایی از نمونه که در ردهٔ C هستند

$Q =$ نسبت تعداد واحدهایی از جامعه که در ردهٔ $\overline{C}$ هستند

$q =$ نسبت تعداد واحدهایی از نمونه که در ردهٔ $\overline{C}$ هستند

بدیهی است $P + Q = ۱$ و $p + q = ۱$.

قضیهٔ ۱۱.۲ در نمونه‌گیری تصادفی سادهٔ بدون جایگذاری، p برآوردکنندهٔ نااریب P است.

برهان. طبق معمول واحدهای جامعه را با $Y_۱$، $Y_۲$، ... ، Y_N و واحدهای نمونه را با $Y_۱$، $Y_۲$، ... ، Y_n نشان می‌دهیم. اگر به واحدهایی که در جامعه یا در نمونه به ردهٔ C تعلق دارند عدد ۱ و به واحدهایی که در ردهٔ $\overline{C}$ هستند عدد ۰ را نسبت دهیم، مقادیر واحدهای جامعه یا نمونه به‌صورت دنباله‌ای از یکها و صفرها درمی‌آیند که تعداد یکها برابر با تعداد واحدهای جامعه در ردهٔ C، یعنی برابر با A است. در نمونه نیز واحدها به‌صورت دنباله‌ای از یکها و صفرها بوده و تعداد یکها برابر با تعداد واحدهایی از نمونه است که در ردهٔ C می‌افتند، یعنی برابر با a است. درواقع از جامعه‌ای که مقادیر متغیر تحت بررسی آن یک و صفرند نمونه‌ای به‌روش بدون جایگذاری انتخاب می‌کنیم لذا باید همهٔ ویژگیها و قضایای مربوط به نمونه‌گیری بدون جایگذاری، در این حالت

خاص هم که مقادیر متغیر تحت بررسی در جامعه یک یا صفرند صادق باشند. همان‌طور که در بالا توضیح دادیم، در جامعه

$$\sum_{i=1}^{N} Y_i = A$$

و در نمونه

$$\sum_{i=1}^{n} Y_i = a$$

لذا

$$\overline{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{A}{N} = P \qquad (۲۱.۲)$$

و

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{a}{n} = p \qquad (۲۲.۲)$$

در مورد نمونهٔ تصادفی سادهٔ بدون جایگذاری دیدیم

$$E(\overline{Y}_n) = \overline{Y}_N$$

با توجه به روابط (۲۱.۲) و (۲۲.۲)

$$E(p) = P$$

بنابراین p، یعنی نسبت تعداد واحدهایی از نمونه که در ردهٔ C هستند برآوردکنندهٔ ناریب P، نسبت تعداد واحدهایی‌از جامعه است که در ردهٔ C قرار دارند و قضیه ثابت می‌شود. □

تذکر. چون تعداد نمونه‌های ممکن $\binom{N}{n}$ است، لذا تعداد مقادیری که p اختیار می‌کند (و ممکن است بعضی از آنها تکراری باشند) برابر $\binom{N}{n}$ است. پس p متغیری تصادفی است که در عمل پس از تهیهٔ یک نمونه، یک مقدار آن را مشاهده می‌کنیم و $\hat{P} = p$.

فرع. برآوردکنندهٔ ناریب تعداد واحدهایی از جامعه که به ردهٔ C تعلق دارند (A) برابر با Np است. زیرا قبلاً دیدیم که $P = \frac{A}{N}$، پس $A = NP$، لذا $\hat{A} = N\hat{P} = Np$. این برآوردکننده، ناریب است، زیرا

$$E(\hat{A}) = E(Np) = NE(p) = NP = A$$

مثال ۱۴.۲ از فهرستی شامل ۲۰۰۰ نام و آدرس، نمونه‌ای تصادفی به‌حجم ۲۰۰ آدرس انتخاب می‌شود. بررسی این نمونه نشان می‌دهد که ۴۰ آدرس صحیح نیستند. نسبت آدرسهای غلط را در کل فهرست برآورد کنید.

اگر P نسبت آدرسهای غلط در جامعه و p نسبت آدرسهای غلط در نمونه باشد، داریم

$$N = ۲۰۰۰ \quad , \quad n = ۲۰۰ \quad , \quad a = ۴۰$$

پس

$$\hat{P} = p = \frac{a}{n} = \frac{۴۰}{۲۰۰} = ۰{,}۲$$

لذا

$$\hat{A} = Np = ۲۰۰۰(۰{,}۲) = ۴۰۰$$

▲

۲.۱۱.۲ محاسبه $S^۲$ و $s^۲$

با نمادهای ۱.۱۱.۲، می‌توان نوشت

$$\sum_{i=۱}^{N} Y_i^۲ = \text{مجموع توان دوم یکها و صفرها در دنبالهٔ جامعه}$$
$$= \text{مجموع توان دوم یکها در دنبالهٔ جامعه}$$
$$= A = NP \qquad \text{(الف)}$$

به‌همین ترتیب

$$\sum_{i=۱}^{n} Y_i^۲ = \text{مجموع توان دوم یکها و صفرها در دنبالهٔ نمونه}$$
$$= \text{مجموع توان دوم یکها در دنبالهٔ نمونه}$$
$$= a = np \qquad \text{(ب)}$$

با توجه به (الف)

$$S^۲ = \frac{۱}{N-۱}\sum_{i=۱}^{N}(Y_i - \overline{Y}_N)^۲ = \frac{۱}{N-۱}\left(\sum_{i=۱}^{N} Y_i^۲ - N\overline{Y}_N^۲\right)$$
$$= \frac{۱}{N-۱}(NP - NP^۲) = \frac{N}{N-۱}P(۱-P) = \frac{N}{N-۱}PQ$$

پس

$$S^{۲} = \frac{N}{N-۱} PQ \tag{۲۳.۲}$$

به‌همین ترتیب، با توجه به (ب)

$$\begin{aligned} s^{۲} &= \frac{۱}{n-۱} \sum_{i=۱}^{n} (Y_i - \overline{Y}_n)^{۲} = \frac{۱}{n-۱} \left(\sum_{i=۱}^{n} Y_i^{۲} - n\overline{Y}_n^{۲} \right) \\ &= \frac{۱}{n-۱}(np - np^{۲}) = \frac{n}{n-۱} p(۱-p) = \frac{n}{n-۱}\, pq \end{aligned}$$

پس

$$s^{۲} = \frac{n}{n-۱} pq \tag{۲۴.۲}$$

قضیهٔ ۱۲.۲ واریانس متغیر تصادفی p به‌صورت زیر است

$$\mathrm{V}(p) = \frac{N-n}{N-۱} \cdot \frac{PQ}{n} \tag{۲۵.۲}$$

برهان. در نمونه‌گیری تصادفی ساده، بدون جایگذاری، داریم

$$\mathrm{V}(\overline{Y}_n) = \left(\frac{۱}{n} - \frac{۱}{N} \right) S^{۲}$$

با توجه به (۲۳.۲) برای نسبتها، و اینکه در این مورد، $\overline{Y}_n = p$، داریم

$$\begin{aligned} \mathrm{V}(p) &= \left(\frac{۱}{n} - \frac{۱}{N} \right) \frac{N}{N-۱} PQ = \frac{N-n}{nN} \cdot \frac{N}{N-۱}\, PQ \\ &= \frac{N-n}{N-۱} \cdot \frac{PQ}{n} \end{aligned}$$

□

در عمل محاسبهٔ واریانس p ممکن نیست، زیرا P در جامعه مجهول است، لذا به کمک نمونه برآوردی برای آن به‌دست می‌آورند.

قضیهٔ ۱۳.۲ برآوردکنندهٔ ناریب $\mathrm{V}(p)$ به‌صورت زیر است

$$\hat{\mathrm{V}}(p) = \hat{\mathrm{V}}(\hat{P}) = \frac{N-n}{n-۱} \frac{pq}{N} \tag{۲۶.۲}$$

برهان. در نمونه‌گیری تصادفی بدون جایگذاری دیدیم

$$\hat{\mathrm{V}}(\overline{Y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \qquad \text{(الف)}$$

با توجه به $\overline{Y}_n = p$ و رابطهٔ (۲۴.۲)، نتیجه می‌شود

$$\begin{aligned}\hat{\mathrm{V}}(p) &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} pq = \frac{N-n}{nN} \cdot \frac{n}{n-1} pq \\ &= \frac{N-n}{n-1} \frac{pq}{N}\end{aligned}$$

نااریب بودن (الف) را قبلاً دیده‌ایم لذا برآوردکنندهٔ (۲۶.۲) نیز نااریب است. □

فرع ۱ . اگر N، حجم جامعه، بزرگ باشد و بتوان $\frac{n}{N}$ را نادیده گرفت، آن‌گاه

$$\hat{\mathrm{V}}(p) = \frac{1 - \frac{n}{N}}{n-1} pq \simeq \frac{pq}{n-1} \qquad (۲۷.۲)$$

گاهی اوقات برآورد واریانس p را فقط $\frac{pq}{n}$ می‌گیرند. البته $\frac{pq}{n}$ برآوردکننده‌ای اریب است.

فرع ۲ . برآوردکنندهٔ نااریب واریانس $\hat{A} = Np$ به‌صورت زیر است

$$\hat{\mathrm{V}}(\hat{A}) = \hat{\mathrm{V}}(Np) = N^2 \hat{\mathrm{V}}(p) = N^2 \frac{N-n}{n-1} \frac{pq}{N} = \frac{N-n}{n-1} \cdot Npq \qquad (۲۸.۲)$$

اگر حجم جامعه بزرگ و $\frac{n}{N}$ قابل اغماض باشد، از (۲۷.۲) نتیجه می‌شود که

$$\hat{\mathrm{V}}(\hat{A}) \simeq \frac{N^2}{n-1} pq \qquad (۲۹.۲)$$

بدیهی است که برآوردکنندهٔ (۲۸.۲) برآوردکننده‌ای نااریب است.

مثال ۱۵.۲ برای داده‌های مثال ۱۴.۲ برآورد واریانس برآوردکنندهٔ نسبت آدرسهای غلط و همچنین برآورد واریانس برآوردکنندهٔ تعداد کل آدرسهای غلط را محاسبه می‌کنیم. یادآوری می‌شود که در مثال ۱۴.۲ .

$$N = ۲۰۰۰ \;,\; n = ۲۰۰ \;,\; a = ۴۰ \;,\; \hat{P} = p = ۰٫۲ \;,\; \hat{A} = ۴۰۰$$

پس،

$$\hat{\mathrm{V}}(\hat{P}) = \frac{۲۰۰۰ - ۲۰۰}{۲۰۰ - ۱} \cdot \frac{(۰٫۲)(۰٫۸)}{۲۰۰۰} \simeq ۰٫۰۰۰۷۲۳$$

اگر $\frac{n}{N} = \frac{۱}{۱۰}$ را قابل اغماض بدانیم

$$\hat{\mathrm{V}}(\hat{P}) \simeq \frac{pq}{n-۱} = \frac{(۰٫۲)(۰٫۸)}{۱۹۹} \simeq ۰٫۰۰۰۸۰۴$$

برآورد واریانس $\hat{A}$ به‌صورت زیر محاسبه می‌شود

$$\hat{\mathrm{V}}(\hat{A}) = \frac{N-n}{n-۱} \cdot Npq = \frac{۲۰۰۰-۲۰۰}{۱۹۹}(۲۰۰۰)(۰٫۲)(۰٫۸) \simeq ۲۸۹۴٫۴۷۲$$

فرع ۳ . از فرمول (۲۶.۲) نتیجه می‌شود که

$$\hat{\sigma}^{۲}(p) = \frac{N-n}{n-۱} \cdot \frac{pq}{N}$$

بنابراین

$$\hat{\sigma}(p) = \sqrt{\frac{N-n}{n-۱} \cdot \frac{pq}{N}} \qquad (۳۰.۲)$$

این برآوردکننده، اریب است. اگر $\frac{n}{N}$ قابل اغماض باشد

$$\hat{\sigma}(p) \simeq \sqrt{\frac{pq}{n-۱}} \qquad (۳۱.۲)$$

به‌همین ترتیب در مورد $\hat{A}$، با توجه به (۲۸.۲) داریم

$$\hat{\mathrm{V}}(\hat{A}) = \hat{\sigma}^{۲}(\hat{A}) = \frac{N-n}{n-۱} \cdot Npq$$

لذا،

$$\hat{\sigma}(\hat{A}) = \sqrt{\frac{N-n}{n-۱} Npq} \qquad (۳۲.۲)$$

این برآوردکننده نیز اریب است. وقتی $\frac{n}{N}$ قابل اغماض باشد، نتیجه می‌شود که

$$\hat{\sigma}(\hat{A}) \simeq N\sqrt{\frac{pq}{n-۱}} \qquad (۳۳.۲)$$

۳.۱۱.۲ حدود اطمینان P

اگر بپذیریم که آمارهٔ p دارای توزیع نرمال است، که برای n بزرگ چنین است، آن‌گاه متغیر تصادفی p از توزیع نرمالی با میانگین P و انحراف معیار $\sigma(p) = \sqrt{\frac{N-n}{N-1} \cdot \frac{PQ}{n}}$ پیروی می‌کند. لذا $Z = \frac{p-P}{\sigma(p)}$ دارای توزیع نرمال استاندارد است. اگر $\phi z(z) = 1 - \frac{\alpha}{2}$، که در آن ϕ تابع توزیع نرمال استاندارد است، آن‌گاه

$$P_r\left(-z < \frac{p-P}{\sigma(p)} < z\right) = 1 - \alpha$$

پس با ضریب اطمینان $1 - \alpha$، داریم

$$-z < \frac{p-P}{\sigma(p)} < z$$

و یا

$$p - z\sigma(p) < P < p + z\sigma(p)$$

بنابراین حدود اطمینان، برای مقدار واقعی P، با ضریب اطمینان $1 - \alpha$ به‌ترتیب $p - z\sigma(p)$ و $p + z\sigma(p)$ هستند. این حدود با تغییر نمونه تغییر می‌کنند. برای یک نمونهٔ مشاهده‌شده، p میانگین نمونه است و z از جدول نرمال استاندارد (متناظر با احتمال تجمعی $1 - \alpha/2$) به‌دست می‌آید. در عمل $\sigma(p)$ معلوم نیست زیرا P و Q مجهول‌اند، بنابراین از برآورد آن، $\hat{\sigma}(p)$ که به‌کمک مشاهدات نمونه از فرمول $\sqrt{\frac{N-n}{n-1} \cdot \frac{pq}{N}}$ به‌دست می‌آید استفاده می‌شود و لذا برآوردهای حدود اطمینان P به‌صورت زیر به‌دست می‌آیند

$$\begin{cases} \text{برآورد حدّ پایین} = p - z\sqrt{\dfrac{N-n}{n-1} \cdot \dfrac{pq}{N}} \\ \text{برآورد حدّ بالا} = p + z\sqrt{\dfrac{N-n}{n-1} \cdot \dfrac{pq}{N}} \end{cases} \qquad (۳۴.۲)$$

اگر $\frac{n}{N}$ قابل اغماض باشد ضریب z در (۳۴.۲) به‌صورت $\sqrt{\frac{pq}{n-1}}$ خواهد بود. در محاسبهٔ حدود اطمینان P، نرمال بودن توزیع متغیر تصادفی p را پذیرفتیم. در آمار ریاضی دیده‌اید که وقتی n بزرگ و p نه خیلی کوچک و نه خیلی بزرگ است توزیع p تقریباً نرمال است. سؤالی که معمولاً مطرح می‌شود این است که حجم نمونه باید به چه بزرگی باشد تا پذیرهٔ نرمال بودن توزیع p معقول تلقی شود. بدیهی است وقتی P به ۱ یا صفر نزدیک باشد توزیع p چاوله است و لذا پذیرهٔ نرمال بودن آن معقول نیست. ککران قاعده‌ای برای انتخاب n به‌منظور درستی تقریبی توزیع نرمال، به‌ازای مقادیر مختلف p ارائه می‌دهد که آن را در جدول زیر آورده‌ایم.

p	تنها وقتی از توزیع نرمال استفاده کنید که n حداقل برابر باشد با:
۰٫۵	۳۰
۰٫۴ یا ۰٫۶	۵۰
۰٫۳ یا ۰٫۷	۸۰
۰٫۲ یا ۰٫۸	۲۰۰
۰٫۱ یا ۰٫۹	۶۰۰
۰٫۰۵ یا ۰٫۹۵	۱۴۰۰

قاعدهٔ بالا به‌گونه‌ای تهیه شده است که احتمال آنکه حد بالای اطمینان کمتر از P باشد بین ۲٫۵ و ۳٫۵ درصد، و احتمال آنکه حد پایین اطمینان بیش از P باشد بین ۲٫۵ و ۱٫۵ درصد است.

مثال ۱۶.۲ قرار است نسبت خانواده‌های با دو فرزند یا بیشتر را در ناحیه‌ای برآورد کنند. در این ناحیه ۸۰۰۰ خانواده ساکن‌اند. نمونه‌ای تصادفی بدون جایگذاری به‌حجم ۱۰۱ انتخاب می‌نمایند و p را برابر با ۰٫۶ به‌دست می‌آورند. بازهٔ اطمینانی به‌طول $6\sigma(p)$ برای P به‌دست آورید.

چون $p = ۰٫۶$، می‌توان مطابق قاعدهٔ مندرج در جدول قبل، پذیرهٔ نرمال بودن توزیع p را معقول دانست. مقدار $\sigma(p)$ را نداریم لذا از برآورد آن استفاده می‌کنیم. با توجه به مقادیر $p = ۰٫۶$، $q = ۰٫۴$، $N = ۸۰۰۰$، $n = ۱۰۱$، داریم

$$\hat{\sigma}(p) = \sqrt{\frac{N-n}{n-۱} \cdot \frac{pq}{N}} = \sqrt{\frac{۸۰۰۰ - ۱۰۱}{۱۰۰} \cdot \frac{(۰٫۶)(۰٫۴)}{۸۰۰۰}} \simeq ۰٫۰۴۸۶$$

لذا بازهٔ اطمینان مورد نظر به‌صورت

$$[۰٫۶ - ۳(۰٫۰۴۸۶) \quad , \quad ۰٫۶ + ۳(۰٫۰۴۸۶)]$$

یا $(۰٫۴۵۴۲, \ ۰٫۷۴۵۸)$ است. ▲

مثال ۱۷.۲ از خانواده‌های ساکن شهر تهران، ۱۰۰ خانواده به تصادف انتخاب کرده‌ایم و دریافته‌ایم که ۶۰ خانواده با پیشنهاد خصوصی کردن دانشگاهها مخالف‌اند. بازهٔ اطمینانی با ضریب اطمینان ۰٫۹۵ برای مخالفان خصوصی کردن دانشگاهها، که در تهران ساکن‌اند، به‌دست آورید.

در این مثال $a = ۶۰$، $n = ۱۰۰$، پس $\hat{P} = p = ۰٫۶$. با توجه به جدول بالا پذیرهٔ نرمال بودن توزیع p، معقول است. چون N، تعداد خانواده‌های ساکن تهران، یعنی حجم جامعه مورد بررسی، بزرگ است می‌توان $\frac{n}{N}$ را قابل اغماض دانست. لذا

$$\hat{\sigma}(p) \simeq \sqrt{\frac{pq}{n-۱}} = \sqrt{\frac{(۰٫۶)(۰٫۴)}{۱۰۰ - ۱}} \simeq ۰٫۰۴۹$$

پس با ضریب اطمینان ۹۵ر۰

$$۰٫۶ - (۱٫۹۶)(۰٫۰۴۹) < P < ۰٫۶ + (۱٫۹۶)(۰٫۰۴۹)$$

یا $۰٫۵۰۴ < P < ۰٫۶۹۶$. ▲

۴.۱۱.۲ حدود اطمینان دقیق P

بازهٔ اطمینانی که برای P ساختیم با این فرض بود که توزیع $p = \hat{P}$ تقریباً نرمال است. حدود بازهٔ اطمینان را می‌توان دقیقتر مبتنی بر توزیع دقیق ابر هندسی تعداد واحدهایی از نمونه که صفت مورد نظر را دارند به‌دست آورد. این روش دقیق از لحاظ توجیه نظری، ساده است ولی از نظر محاسباتی پیچیده است:

فرض کنید طبق معمول $a = \sum_{i=۱}^{n} Y_i$ تعداد واحدهایی در نمونهٔ تصادفی به‌حجم n باشد که برابر با یک‌اند، یعنی صفت مورد نظر را دارند. برای محاسبهٔ حدود دقیق بازهٔ اطمینان، مدل آوند را در نظر می‌گیریم. تصور کنید که آوندی دارای τ مهرهٔ قرمز (صفت مورد نظر) و $N - \tau$ مهرهٔ سفید باشد. درواقع محتوای آوند همان جامعه به‌حجم N است. نمونه‌ای تصادفی بدون جایگذاری و به‌حجم n از آوند اختیار می‌کنیم. فرض کنیم، X متغیری تصادفی باشد که تعداد مهره‌های قرمز نمونه را نشان می‌دهد. با توجه به اینکه τ مهره قرمز در آوند داریم، احتمال اینکه تعداد مهره‌های قرمز در نمونهٔ تصادفی به‌حجم n از آوند، برابر j باشد، با توجه به توزیع ابر هندسی که در آمار ریاضی دیده‌ایم برابر است با

$$\Pr(X = j|\tau) = \binom{\tau}{j}\binom{N-\tau}{n-j}\Big/\binom{N}{n}$$
$$\max(n + \tau - N,\ ۰) \le j \le \min(\tau,\ n)$$

هدف ما درواقع برآورد τ یعنی تعداد مهره‌های قرمز آوند یا درواقع برآورد تعداد واحدهای با صفت مورد نظر در جامعه است. لاجرم باید برای τ حدود اطمینانی تهیه کنیم. اگر حدود اطمینان $۱۰۰(۱ - \alpha)$ درصد مورد نظر باشد و اگر حد بالایی را با τ_U نشان دهیم، این حد برابر تعداد مهره‌های قرمز آوند است که برای به‌دست آوردن a مهرهٔ قرمز یا کمتر در نمونه، احتمال $\alpha_۱$ را تخصیص می‌دهد. عدد $\alpha_۱$ تقریباً برابر با نصف α است. پس τ_U در رابطهٔ زیر صدق می‌نماید

$$\Pr(X \le a|\tau_U) = \sum_{i=۰}^{a}\binom{\tau_U}{i}\binom{N-\tau_U}{n-i}\Big/\binom{N}{n} = \alpha_۱$$

اگر حد پایین τ را با τ_L نشان دهیم، τ_L برابر تعداد مهره‌های قرمز آوند است که برای به‌دست آوردن a مهرهٔ قرمز یا بیشتر احتمال $\alpha_۲$ را اختصاص می‌دهد، که در آن $\alpha_۲$ تقریباً برابر با نصف

α است. پس τ_L در رابطهٔ زیر صدق می‌کند

$$\Pr(X \geq a|\tau_L) = \sum_{i=a}^{n} \binom{T_L}{i}\binom{N-\tau_L}{n-i} \bigg/ \binom{N}{n} = \alpha_۲$$

به‌کمک توزیع ابرهندسی و دو رابطهٔ بالا باید τ_L و τ_U را محاسبه کرد، که البته با پیچیدگی محاسباتی همراه است. بازهٔ اطمینانی که برای τ بدین طریق به‌دست می‌آید کوتاهتر و یا درواقع دقیقتر از بازهٔ اطمینان معمول است. بدیهی است که حدود اطمینان برای P به‌ترتیب $P_L = \frac{\tau_L}{N}$ و $P_U = \frac{\tau_U}{N}$ هستند. اگر $\alpha_۱$ و $\alpha_۲$ از قبل انتخاب شوند، τ_U را باید به‌صورت بزرگترین عدد صحیحی انتخاب کرد که داشته باشیم

$$\Pr(X \leq a|\tau_U) > \alpha_۱$$

و T_L را باید به‌صورت کوچکترین عدد صحیحی انتخاب کرد که داشته باشیم

$$\Pr(X \geq a|\tau_L) > \alpha_۲$$

این شیوه، وجود یک پوشش احتمالاتی حداقل $۱ - \alpha_۱ - \alpha_۲$ را تضمین می‌کند. شرط عدد صحیح بودن τ، ممکن است امکان دستیابی به یک ضریب اطمینان دقیق $۱ - \alpha$ را ضعیف کند.

۱۲.۲ تعیین برآورد نسبتها در جامعه‌ای با چند رده

اگر N واحد جامعه به‌جای دو رده در $K>۲$ رده قرار گیرند تعداد واحدهای ردهٔ i را A_i و نسبت تعداد این واحدها را در جامعه با P_i نشان می‌دهیم. اگر بخواهیم P_i، $i = ۱, \ldots, K$، را برآورد کنیم نمونه‌ای تصادفی بدون جایگذاری به‌حجم n از جامعه می‌گیریم و تعداد واحدهایی از نمونه را که در ردهٔ iام قرار می‌گیرند با a_i، $i = ۱, \ldots, K$ مشخص می‌کنیم. بدیهی است

$$p_i = \frac{a_i}{n} \quad , \quad i = ۱, \ldots, n$$

که برابر با نسبت واحدهایی از نمونه است که در ردهٔ iام می‌افتند. اینک همهٔ $N - A_i$ واحد دیگر جامعه را در رده‌ای دیگر فرض می‌کنیم و روشن است که همهٔ $n - a_i$ واحد نمونه هم در این رده قرار دارند. بدین صورت جامعه مثل قبل تنها به دو رده، رده‌بندی می‌شود، و آنچه برای تعیین نسبتها در دو رده گفتیم در اینجا صادق است. بنابراین، داریم

$$\hat{P}_i = p_i = \frac{a_i}{n}$$

به‌همین ترتیب می‌توان برآورد نسبت واحدهای جامعه را در هر یک از K رده معین کرد. واریانس و برآورد واریانس برآوردکنندهٔ نسبت هر رده نیز به روش دو رده‌ای کردن جامعه و نمونه به‌صورت قبل حساب می‌شوند.

مثال ۱۸.۲ از ۵۰۰۰ ساکنین شهرکی، نمونه‌ای به‌حجم ۱۰۰ به‌روش تصادفی بدون جایگذاری انتخاب کرده‌اند. پس از آزمایش ۴۱ نفر دارای گروه خونی A، ۴۹ نفر دارای گروه خونی B و ۱۰ نفر دارای گروه خونی O بودند. نسبت هر گروه خونی را برای ساکنین شهرک برآورد کنید و واریانس برآوردکنندهٔ هر نسبت را برآورد نمایید.

با توجه به نمادگذاریها

$$N = ۵۰۰۰, \quad n = ۱۰۰, \quad a_۱ = ۴۱, \quad a_۲ = ۴۹ \quad a_۳ = ۱۰$$

اگر نسبتها را در جامعه با $P_۱$، $P_۲$، و $P_۳$، و در نمونه به‌ترتیب با $p_۱$، $p_۲$، $p_۳$ نشان دهیم

$$p_۱ = \hat{P}_۱ = \frac{۴۱}{۱۰۰} \quad p_۲ = \hat{P}_۲ = \frac{۴۹}{۱۰۰} \quad p_۳ = \hat{P}_۳ = \frac{۱۰}{۱۰۰}$$

برای محاسبهٔ برآورد واریانس $p_۱$، جامعه را به دو رده، با گروه خونی A و غیر A، رده‌بندی می‌کنیم. برای این دو رده

$$p_۱ = ۰٫۴۱ \quad q_۱ = ۰٫۵۹$$

پس

$$\hat{\mathrm{V}}(p_۱) = \frac{N-n}{n-۱}\,\frac{p_۱q_۱}{N} = \frac{۵۰۰۰ - ۱۰۰}{۹۹}\;\frac{(۰٫۴۱)(۰٫۵۹)}{۵۰۰۰} \simeq ۰٫۰۰۲۳$$

به‌همین ترتیب ۴۹ نفر در نمونه دارای گروه خونی B و ۵۱ نفر دارای گروه خونی غیر B هستند، لذا برای دو ردهٔ اخیر

$$p_۲ = ۰٫۴۹ \quad q_۲ = ۰٫۵۱$$

پس

$$\hat{\mathrm{V}}(p_۲) = \frac{۵۰۰۰ - ۱۰۰}{۹۹} \cdot \frac{(۰٫۴۹)(۰٫۵۱)}{۵۰۰۰} \simeq ۰٫۰۰۲۴$$

و بالاخره ۱۰ نفر از نمونه دارای گروه خونی O و بقیه دارای گروه خونی غیر O هستند، و برای این دو رده

$$p_۳ = ۰٫۱ \quad q_۳ = ۰٫۹$$

پس

$$\hat{\mathrm{V}}(p_۳) = \frac{۵۰۰۰ - ۱۰۰}{۹۹} \cdot \frac{(۰٫۹)(۰٫۱)}{۵۰۰۰} \simeq ۰٫۰۰۰۸۹$$

▲

۱۳.۲ برآورد میانگین و نسبت در یک زیرجامعه

در یک بررسی نمونه‌ای برای برآورد نسبت رأی‌دهندگانی که ثبت نام کرده‌اند و قصد دارند به نفع نامزدی خاص رأی دهند نمونه‌ای تصادفی به‌حجم n از ناحیهٔ رأی‌گیری انتخاب می‌کنیم. از افرادی که در نمونه هستند $n_۱$ نفر جز رأی‌دهندگان ثبت نام کرده هستند، و از این $n_۱$ نفر $a_۱$ نفر قصد دارند به نامزد خاص رأی دهند. نسبت رأی‌دهندگان ثبت نام کردهٔ موافق با نامزد مورد نظر، با نسبت نمونه‌ای $\hat{p}_۱ = \frac{a_۱}{n_۱}$ برآورد می‌شود. در این فرایند، نسبتی که باید برآورد شود نسبتی از کل جامعه نیست ، بلکه نسبتی از زیرجامعه (یا حوزهٔ بررسی) است. این زیرجامعه، زیر جامعه رأی‌دهندگانی است که برای رأی دادن ثبت نام کرده‌اند. در رابطهٔ $\hat{p}_۱ = \frac{a_۱}{n_۱}$ نه تنها $a_۱$ متغیری تصادفی است، بلکه $n_۱$ نیز متغیری تصادفی است. وضعیتی مشابه وقتی رخ می‌دهد که نمونه‌ای تصادفی از خانوارها را برای برآورد کردن میانگین یا مجموع هزینه‌ها انتخاب می‌کنیم، اما برآوردها را برای کل جامعه نمی‌خواهیم بلکه برای زیرجامعه‌هایی مبتنی بر حجم خانواده، ناحیهٔ جغرافیایی یا عاملهای دیگر تهیه می‌کنیم.

۱.۱۳.۲ میانگین زیرجامعه و برآورد آن

فرض کنید از N واحد جامعه، N_K واحد به زیرجامعهٔ Kام متعلق باشند. مقدار متغیر واحد iام از زیرجامعهٔ Kام را با Y_{Ki} نشان می‌دهیم. مجموع واحدهای زیرجامعهٔ Kام، $\tau_K = \sum_{i=۱}^{N_K} Y_{Ki}$ و میانگین آن $\mu_K = \frac{\tau_K}{N_K}$. از N واحد کل جامعه، نمونه‌ای تصادفی به‌حجم n انتخاب می‌کنیم و متوجه می‌شویم که n_K واحد آن در زیرجامعهٔ Kام است (n_K متغیری تصادفی است). میانگین نمونه‌ای این n_K واحد برابر است با

$$\overline{Y}_K = (۱/n_K) \sum_{i=۱}^{n_K} Y_{Ki}$$

آمارهٔ $\overline{Y}_K$ برآوردکنندهٔ نااریب میانگین زیرجامعهٔ Kام یعنی برآوردکنندهٔ μ_K است. برای اثبات این ادعا، از روابط (۲۳.۱) و (۲۴.۱) کمک می‌گیریم. با فرض معلوم بودن n_K، هر ترکیب ممکن n_K واحدی از N_K واحد زیرجامعهٔ Kام، شانسی یکسان برای ظاهرشدن در نمونه را دارد. پس، به‌شرط n_K، $\overline{Y}_K$ به‌عنوان میانگین نمونه‌ای نمونهٔ تصادفی سادهٔ به حجم n_K از N_K واحد به‌حساب می‌آید و باید داشته باشیم

$$E(\overline{Y}_K | n_K) = \mu_K$$

ولی در حقیقت n_K خود متغیری تصادفی است لذا امید غیرشرطی $\overline{Y}_K$ به‌صورت زیر محاسبه می‌شود

$$E(\overline{Y}_K) = E[E(\overline{Y}_K|n_K)] = \mu_K$$

بنابراین $\overline{Y}_K$ برآوردکنندهٔ ناارِیب μ_K است.

۲.۱۳.۲ واریانس برآوردکنندهٔ میانگین زیرجامعه

واریانس $\overline{Y}_K$ از فرمول واریانس حاشیه‌ای (۲۴.۱) به‌صورت زیر به‌دست می‌آید.

$$\mathrm{V}(\overline{Y}_K) = E[\mathrm{V}(\overline{Y}_K|n_K)] + \mathrm{V}[E(\overline{Y}_K|n_K)] \qquad \text{(الف)}$$

در این رابطه، جملهٔ دوم طرف دوم برابر با صفر است، زیرا در آن امید شرطی مقداری ثابت است. لذا با توجه به واریانس میانگین در نمونه‌گیری تصادفی ساده،

$$\mathrm{V}(\overline{Y}_K|n_K) = \left(\frac{۱}{n_K} - \frac{۱}{N_K}\right) S_K^۲$$

که $S_K^۲$ تغییرات واحدهای زیرجامعهٔ Kام است، یعنی

$$S_K^۲ = \frac{۱}{N_K - ۱}\sum_{i=۱}^{N_K}(Y_{Ki} - \mu_K)^۲$$

لذا واریانس حاشیه‌ای $\overline{Y}_K$، بنابر (الف) و توضیحات بالا به‌صورت زیر است

$$\mathrm{V}(\overline{Y}_K) = E\left[\left(\frac{۱}{n_K} - \frac{۱}{N_K}\right) S_K^۲\right]$$

چون $S_K^۲$ و N_K مقادیری ثابت‌اند و n_K متغیری تصادفی است، پس

$$\mathrm{V}(\overline{Y}_K) = S_K^۲\left[E\left(\frac{۱}{n_k} - \frac{۱}{N_K}\right)\right]$$

تذکر. امید متغیر تصادفی n_K، برابر با $n \cdot \frac{N_K}{N}$ است. زیرا n_K تعداد واحدهای زیرجامعهٔ Kام در n واحد نمونه‌ای از کل N واحد جامعه است. درواقع n_K تعداد پیروزیها در توزیعی دوجمله‌ای با n امتحان و شانس پیروزی $\frac{N_K}{N}$ است. پس $E(n_K) = n \cdot \frac{N_K}{N}$. طرحی با حجم نمونه‌ای ثابت $n\frac{N_K}{N}$ کمترین واریانس را به‌دست می‌دهد زیرا براساس نابرابری ینسن که در احتمال داریم، $E\left(\frac{۱}{n_K}\right)$ بزرگتر از $N/(nN_K)$ است. در حالت کلی $E\left(\frac{۱}{n_K}\right)$ در دست نیست، ولی گاهی آن را از تقریب $E\left(\frac{۱}{n_K}\right) \simeq \frac{۱}{E(n_K)}$ به‌دست می‌آورند.

۳.۱۳.۲ برآورد واریانس $\overline{Y}_K$

برآوردکننده‌ای برای $\mathrm{V}(\overline{Y}_K)$ به‌صورت زیر است:

$$\hat{\mathrm{V}}(\overline{Y}_K) = \left(\frac{1}{n_K} - \frac{1}{N_K}\right) s_K^2$$

که در آن

$$s_K^2 = \frac{1}{n_K - 1} \sum_{i=1}^{n_K} (Y_{Ki} - \overline{Y}_K)^2$$

آمارهٔ s_K^2، برآورد نااریب واریانس شرطی $\overline{Y}_K$ به‌شرط n_K است. یک بازهٔ اطمینان $100(1-\alpha)$ درصد برای μ_K با حدود $\overline{Y}_K \pm t\sqrt{\hat{\mathrm{V}}(\overline{Y}_K)}$ تهیه می‌شود که در آنها t نقطهٔ صدک $1-\frac{\alpha}{2}$ توزیع t با $n_K - 1$ درجهٔ آزادی است. چون شیوهٔ بازهٔ اطمینان، پوشش احتمال تقریبی مطلوب را به‌صورت شرطی برای حجم نمونه‌ای n_k داراست، پوشش احتمال غیرشرطی را نیز به‌طریق اولی دارا خواهد بود.

اگر N_K مجهول باشد که غالباً چنین است، به‌جای عامل تصحیح متناهی $(N_K - n_K)/N_K$ می‌توان امید آن، $(N-n)/N$ را قرار داد.

۴.۱۳.۲ برآورد نسبت در یک زیرجامعه

نسبت در زیرجامعه، حالتی خاص از میانگین زیرجامعه است، زیرا در این حالت مقدار Y_{Ki} برای واحدی از زیرجامعهٔ Kام که مشخصهٔ مورد نظر را دارد برابر با ۱ است و در غیر این صورت برابر ۰ است. اگر تعداد واحدهای با مشخصهٔ مورد نظر را در زیرجامعهٔ Kام با A_K نشان دهیم وقتی نمونه‌ای به‌حجم n از کل N واحد جامعه به‌تصادف بگیریم، n_K واحد آن متعلق به زیرجامعهٔ Kام است و از این n_K واحد، a_K واحد دارای صفت مورد نظر است. در حالت کلی n_K و a_K متغیرهایی تصادفی‌اند. با توجه به حالت قبل و اینکه حالت مورد بحث، حالتی خاص از آن است، $\hat{P}_K = \frac{a_K}{n_K}$ که در آن، P_K نسبت واحدهای با مشخصهٔ مورد نظر در زیرجامعهٔ Kام است. $\hat{P}_K$ به‌عنوان میانگین واحدهای ۱ و صفر نمونه‌ای زیرجامعه، برای $P = \frac{A_K}{N_K}$ برآوردکننده‌ای نااریب است. به‌سادگی نتیجه می‌شود که

$$\hat{\mathrm{V}}(\hat{P}_K) = \left(\frac{N_K - n_K}{N_K}\right) \frac{\hat{P}_K(1-\hat{P}_K)}{n_K - 1}$$

وقتی N_K مجهول است می‌توان به‌جای $(N_K - n_K)/N_K$ مقدار $(N-n)/N$ را به‌صورت تقریب منظور کرد.

۱۴.۲ برآورد مجموع واحدهای یک زیرجامعه

اگر در جامعهٔ به‌حجم N، حجم یک زیرجامعه، N_K، معلوم باشد و اگر T_K مجموع کل این زیرجامعه فرض شود، می‌توان T_K را به‌صورت نااریب با $N_K\overline{Y}_K$ برآورد کرد. آمارهٔ $\overline{Y}_K$ میانگین بخشی از نمونه‌ای تصادفی به‌حجم n است که به‌تصادف از کل جامعه گرفته‌ایم و به زیرجامعه تعلق داشته، حجم آن n_K است.

$$\mathrm{V}(\hat{T}_K) = \mathrm{V}(N_K\overline{Y}_K) = N_K^{۲}\mathrm{V}(\overline{Y}_K)$$

و برآورد این واریانس، یعنی

$$\hat{\mathrm{V}}(N_K\overline{Y}_K) = N_K^{۲}\hat{\mathrm{V}}(\overline{Y}_K)$$

را می‌توان مورد استفاده قرار داد. اگر N_K مجهول باشد برآوردکنندهٔ نااریب

$$\hat{T}_K = \frac{N}{n}\sum_{i=۱}^{n_K} Y_{Ki}$$

را می‌توان به‌کار برد. برای اینکه نشان دهیم این برآوردکننده نااریب است متغیر جدید Y'را بدین صورت تعریف می‌کنیم که هر واحد Yی نمونه به‌حجم n اگر به زیرجامعه تعلق داشته باشد برابر با خودش و اگر به این زیرجامعه متعلق نباشد برابر ۰ است. برای متغیر جدید، مجموع کل جامعه برابر است با $\sum_{i=۱}^{N} Y'_i = T_K$. واریانس این متغیر جدید را $\sigma'^{۲}$ می‌گیریم. با نمونهٔ به‌حجم n که از جامعهٔ اصلی گرفته‌ایم. میانگین نمونه‌ای متغیر جدید عبارت است از

$$\overline{Y}' = \sum_{i=۱}^{n} \frac{Y'_i}{n}$$

برآوردکنندهٔ T_K را می‌توان به‌صورت

$$\hat{T}_K = N\overline{Y}'$$

نوشت. بدیهی است $\hat{T}_K$ برآوردکنندهٔ نااریب T_K است. پس با توجه به دو رابطه اخیر

$$\hat{T}_K = \frac{N}{n}\sum_{i=۱}^{n} Y'_i$$

اما از n واحد Y'_i، تعداد $n - n_K$ تا برابر با صفر و n_K تا برابر واحدهای نمونهٔ اصلی‌اند که به زیرجامعه تعلق دارند، یعنی

$$\hat{T}_K = \frac{N}{n}\sum_{i=۱}^{n_K} Y_{Ki}$$

۱۵.۲ برآورد حجم نمونه

در طراحی بررسی نمونه‌ای، اخذ تصمیم دربارهٔ حجم نمونه، از لحاظ تأمین میزان دقت نتایج نمونه‌گیری و صرفه‌جویی در مقدار وقت و هزینه، از اهمیتی خاص برخوردار است. بدیهی است که بزرگ بودن حجم نمونه موجب صرف هزینه و وقت زیاد، و کوچک بودن حجم نمونه موجب عدم دقت کافی برآوردها می‌شود. همواره سعی بر آن است که درچارچوب اطلاعات موجود و با توجه به وقت و هزینهٔ ممکن و دقت لازم، بهترین حجم ممکن نمونه را انتخاب کنند. در نظریهٔ نمونه‌گیری مراحل اصلی انتخاب حجم نمونه به‌شرح زیرند:

۱. باید دربارهٔ انتظاری که از استنباط نمونه‌ای داریم حکمی روشن تهیه کنیم. این حکم ممکن است برحسب حدود خطا باشد و یا ممکن است برحسب تصمیمی باشد که قرار است بعد از تعیین نتایج نمونه اخذ شود و یا برحسب عملی باشد که باید صورت گیرد. مسؤولیت تهیهٔ حکم به‌عهدهٔ افرادی است که نتایج حاصل از نمونه را به‌کار می‌برند. بدیهی است خواسته‌های این افراد باید از صورت کیفی به‌صورت کمّی تبدیل و با عدد بیان شوند.

۲. باید با توجه به محتوای حکم ۱، تقریبی که از نتایج انتظار داریم، و با در نظرگرفتن نوع نمونه‌گیری، معادله‌ای تشکیل دهیم که شامل n، یعنی حجم نمونه باشد. یکی از ویژگیهای نمونه‌گیری احتمالاتی، این است که ما را به تشکیل چنین معادله‌ای قادر می‌سازد. در این معادله می‌توان شرایطی نظیر هزینهٔ نمونه‌گیری را نیز منظور داشت.

۳. در این معادله علاوه‌بر مجهول n، پارامترهای مجهول دیگری که معمولاً همان پارامترهای جامعه‌اند وجود دارند. این پارامترها را باید قبل از انجام نمونه‌گیری اصلی که پس از تعیین n صورت می‌گیرد، به‌وسیلهٔ نمونه‌ای مقدماتی برآورد کرده و در معادله قرار داد. بدیهی است با گنجاندن این برآورد در معادله، مقداری که برای n به‌دست می‌آید برآوردی از n واقعی است.

۴. اگر جامعه به رده‌هایی تقسیم شده باشد و حدود خطا یا احکام مربوط به داده‌ها متفاوت باشند، باید برای هر رده حجم نمونه را جداگانه برآورد و سپس حجم کل نمونه را معین کرد.

۵. غالباً در یک بررسی نمونه‌ای بیش از یک مشخصه را مطالعه می‌کنند. اگر درجهٔ دقت مطلوب برای مشخصه‌ها متفاوت باشد برای هر مشخصه یک n به‌دست می‌آید. در این صورت باید روشی برای تلفیق این nها در نظرگرفت. معمولاً n کلی را برابر با بزرگترین n حاصل در دنبالهٔ nها اختیار می‌کنند. پس از این مقدمه، چندین روش برآورد حجم نمونه را که بیشتر در کاربردها از آنها استفاده می‌کنیم مطرح می‌نماییم.

۱۶.۲ محاسبهٔ برآورد حجم نمونه با توجه به خطای نسبی برآورد

در هر نمونه، $\overline{Y}_n$ در حالت کلی با $\overline{Y}_N$ متفاوت است و وقتی آن را به‌عنوان برآورد $\overline{Y}_N$ اختیار می‌کنیم مقدار خطای برآورد برابر با $\overline{Y}_n - \overline{Y}_N$ است. معمولاً $\frac{\overline{Y}_n-\overline{Y}_N}{\overline{Y}_N}$ را خطای نسبی می‌نامند. چون $\overline{Y}_n$ متغیری تصادفی است، بنابراین خطای نسبی نیز متغیری تصادفی است و در یک

نمونه‌گیری تصادفی بدون جایگذاری، تعداد $\binom{N}{n}$ مقدار می‌پذیرد که ممکن است بعضی از آنها تکراری باشند. گاهی حکمی که برای تعیین n در نظر می‌گیریم این است که n را به‌گونه‌ای بیابیم که قدر مطلق خطای نسبی با احتمال $\alpha - ۱$ از مقدار مفروض r کوچکتر باشد. یعنی

$$\Pr\left(\left|\frac{\overline{Y}_n - \overline{Y}_N}{\overline{Y}_N}\right| < r\right) = ۱ - \alpha \qquad (۳۵.۲)$$

و یا

$$\Pr\left(\left|\frac{\overline{Y}_n - \overline{Y}_N}{\overline{Y}_N}\right| \geq r\right) = \alpha$$

که در آنها α احتمالی بسیار کوچک است و با توجه به‌دقت مورد نیاز، از قبل به‌وسیلهٔ مصرف‌کنندهٔ اطلاعات نمونه، یا به‌وسیلهٔ آماردان تعیین می‌شود. از این حکم احتمالاتی وقتی می‌توان استفاده کرد که توزیع $\overline{Y}_n$ نرمال یا تقریباً نرمال باشد. با این فرض (۳۵.۲) را به‌صورت زیر می‌نویسم

$$\Pr\left(\left|\frac{\overline{Y}_n - \overline{Y}_N}{\overline{Y}_N}\right| < r\right) = \Pr(|\overline{Y}_n - \overline{Y}_N| < r\overline{Y}_N)\ , \qquad \overline{Y}_N > ۰$$

یا

$$\Pr(-r\overline{Y}_N < \overline{Y}_n - \overline{Y}_N < r\overline{Y}_N) = ۱ - \alpha$$

با توجه به $\sigma(\overline{Y}_n) = S\sqrt{\frac{N-n}{nN}}$،

$$\Pr\left(\frac{-r\overline{Y}_N}{\sigma(\overline{Y}_n)} < \frac{\overline{Y}_n - \overline{Y}_N}{\sigma(\overline{Y}_n)} < \frac{r\overline{Y}_N}{\sigma(\overline{Y}_n)}\right) = ۱ - \alpha$$

ولی $Z = \frac{\overline{Y}_n - \overline{Y}_N}{\sigma(\overline{Y}_n)}$ دارای توزیع نرمال استاندارد است و اگر قرار دهیم $\phi_Z(z) = ۱ - \alpha/۲$ آن‌گاه

$$\frac{r\overline{Y}_N}{\sigma(\overline{Y}_n)} = z$$

یا

$$r\overline{Y}_N = zS\sqrt{\frac{N-n}{nN}}$$

که اگر پارامترهای S و $\overline{Y}_N$ معلوم باشند از حل این معادله می‌توان n را به‌دست آورد. پس از حل معادله، داریم

$$n = \left(\frac{zS}{r\overline{Y}_N}\right)^۲ \Big/ \left[۱ + \frac{۱}{N}\left(\frac{zS}{r\overline{Y}_N}\right)^۲\right] \qquad (۳۶.۲)$$

در این رابطه، r، حد بالای خطای نسبی است که از قبل تعیین می‌شود، S و $\overline{Y}_N$ پارامترهای جامعه‌اند که اگر معلوم نباشند باید از روی نمونه‌ای مقدماتی برآورد آنها را به‌دست آورد، و z، طول نقطهٔ متناظر با احتمال تجمعی $\alpha - ۱$ یا صدک $(\alpha - ۱)$ام توزیع نرمال استاندارد است. در رابطهٔ (۳۶.۲) مقدار $\frac{S}{\overline{Y}_N}$ تقریباً همان ضریب تغییرات جامعه است. اگر N بزرگ باشد می‌توان (۳۶.۲) را به‌صورت زیر نوشت. اگر قرار دهیم

$$n_۰ = \left(\frac{z}{r}\cdot\frac{S}{\overline{Y}_N}\right)^۲ \qquad (۳۷.۲)$$

از ادغام دو رابطهٔ (۳۶.۲) و (۳۷.۲) نتیجه می‌شود که

$$n = \frac{n_۰}{۱ + n_۰/N} \qquad (۳۸.۲)$$

وقتی $S^۲$ و $\overline{Y}_N$ مجهول‌اند (که معمولاً چنین‌اند) از روی نمونهٔ مقدماتی به‌حجم $n_۱$، $\hat{S}^۲$ و $\hat{\overline{Y}}_N$ یعنی به‌ترتیب واریانس و میانگین نمونه را به‌دست می‌آوریم و در (۳۶.۲) منظور می‌کنیم تا برآوردی برای n حاصل شود

$$\hat{n} = \left(\frac{z}{r}\cdot\frac{s}{\overline{Y}_n}\right)^۲ \Big/ \left[۱ + \frac{۱}{N}\left(\frac{z}{r}\cdot\frac{s}{\overline{Y}_n}\right)^۲\right] \qquad (۳۹.۲)$$

مثال ۱۹.۲ تعداد دانشجویانی که در دانشگاههای تهران تحصیل می‌کنند و اجارهٔ مسکن می‌پردازند ۴۵۰۰۰ نفر است. می‌خواهیم میانگین اجارهٔ مسکنی را که هر دانشجو می‌پردازد برآورد کنیم. مایلیم با احتمال ۹۵ر۰ قدر مطلق خطای برآورد از ۰۴ر۰ میانگین واقعی کمتر باشد. برای محاسبهٔ چنین برآوردی حجم نمونه را برآورد کنید. مجریان نمونه‌گیری به‌کمک نمونه‌ای تصادفی و بدون جایگذاری و به‌حجم ۵۰ برآوردهای مقدماتی ۱۵۰۰۰ تومان و ۵۰۰۰ تومان را برای $\overline{Y}_N$ و S جامعه به‌دست آورده‌اند.

$\overline{Y}_n$ را میانگین نمونه‌ای می‌گیریم که می‌خواهیم حجم آن را با توجه به شرط مذکور معین کنیم. بنابر شرط، باید

$$\Pr(|\overline{Y}_n - \overline{Y}_N| < ۰٫۰۴\overline{Y}_N) = ۰٫۹۵$$

و یا

$$\Pr\left(\left|\frac{\overline{Y}_n - \overline{Y}_N}{\overline{Y}_N}\right| < ۰{,}۰۴\right) = ۰{,}۹۵$$

که با (۳۵.۲) هماهنگ است. لذا $r = ۰{,}۰۴$. با توجه به $\hat{S} = ۵۰۰۰$ و $\hat{\overline{Y}}_N = ۱۵۰۰۰$، اگر به‌طور تقریبی z را برابر با ۲ بگیریم، داریم

$$\begin{aligned}\hat{n} &= \left(\frac{۲}{۰{,}۰۴}\cdot\frac{۵۰۰۰}{۱۵۰۰۰}\right)^{۲} \Big/ \left[۱ + \frac{۱}{۴۵۰۰۰}\left(\frac{۲}{۰{,}۰۴}\cdot\frac{۵۰۰۰}{۱۵۰۰۰}\right)^{۲}\right] \\ &= \left(\frac{۱}{۰{,}۰۰۶}\right)^{۲} \Big/ \left[۱ + \frac{۱}{۴۵۰۰۰}\left(\frac{۱}{۰{,}۰۰۶}\right)^{۲}\right] \simeq ۲۷۷\end{aligned}$$

▲

۱۷.۲ محاسبهٔ برآورد حجم نمونه برحسب طول معلوم بازهٔ اطمینان

گاهی حجم نمونه را به‌قسمی تعیین می‌کنند که بازهٔ اطمینان جامعه با ضریب اطمینان معلوم $۱ - \alpha$ به‌طول مفروض $۲l$ باشد. اگر f، کسر نمونه‌گیری را بدانیم و شرط نرمال بودن توزیع $\overline{Y}_n$ را بپذیریم بازهٔ اطمینان به‌صورت زیر است

$$\left(\overline{Y}_n - S\sqrt{\frac{۱-f}{n}}\,z \quad , \quad \overline{Y}_n + S\sqrt{\frac{۱-f}{n}}\,z\right).$$

که در آن z از $\phi_Z(z) = ۱ - \alpha/۲$ به‌دست می‌آید. ϕ تابع توزیع احتمال متغیر تصادفی نرمال استاندارد است. مایلیم طول این بازه برابر با $۲l$ باشد، یعنی

$$\overline{Y}_n + S\sqrt{\frac{۱-f}{n}}\,z - \left(\overline{Y}_n - S\sqrt{\frac{۱-f}{n}}z\right) = ۲l$$

یا

$$S\sqrt{\frac{۱-f}{n}}\,z = l$$

که نتیجه می‌دهد

$$n = \frac{S^۲ z^۲}{l^۲}(۱-f)$$

در عمل $S^{۲}$ مجهول است. به‌کمک نمونه‌ای مقدماتی آن را با $s^{۲}$ی نمونهٔ مقدماتی برآورد می‌کنیم پس

$$\hat{n} = \frac{s^{۲}z^{۲}}{l^{۲}}(۱ - f) \qquad (۴۰.۲)$$

اگر f کوچکتر از ۰٫۰۵ باشد معمولاً از آن صرف‌نظر می‌کنند. می‌بینید که هر چه طول بازهٔ اطمینان بیشتر باشد $\hat{n}$ کوچکتر است.

مثال ۲۰.۲ قرار است برای برآورد متوسط قطر درختان ناحیه‌ای از جنگل، نمونه‌گیری تصادفی بدون جایگذاری انجام شود. مقدار $s^{۲}$ در نمونه‌ای مقدماتی برابر با ۳۰۰ شده است. واحد اندازه‌گیری سانتیمتر است. اگر بخواهیم بازهٔ اطمینان میانگین واقعی قطر درختان این ناحیه، با ضریب اطمینان ۰٫۹۵ برابر با ۴ سانتیمتر باشد و کسر نمونه‌گیری ۰٫۰۵ فرض شود، حجم نمونه‌ای را که باید تهیه کنیم برآورد کنید.

با توجه به جدول توزیع نرمال $z_{۰٫۹۷۵} \simeq ۲$. براساس داده‌ها، $l = ۲$، $s^{۲} = ۳۰۰$، $f = ۰٫۰۵$، پس بنابر (۴۰.۲)

$$\hat{n} = \frac{s^{۲}z^{۲}}{l^{۲}}(۱ - f) \simeq \frac{۳۰۰(۴)}{۴}(۱ - ۰٫۰۵) = ۲۸۵$$

▲

۱۸.۲ برآورد حجم نمونه در نمونه‌گیری برای برآورد نسبتها

فرض کنیم واحدهای جامعه در دو رده قرار دارند. اگر p را برآورد نسبت واحدهایی از جامعه بگیریم که در ردهٔ C هستند، مایلیم n، حجم نمونه، را به‌قسمی تعیین کنیم که با احتمال کوچک α، مقدار قدر مطلق خطا، $|p - P|$، برابر با d یا بزرگتر از آن باشد، یعنی

$$\Pr(|p - P| \geq d) = \alpha \qquad (\text{الف})$$

که در آن، P نسبت واقعی واحدهایی از جامعه است که در ردهٔ C هستند و d عددی است که از پیش تعیین شده است. اگر بپذیریم که توزیع p، تقریباً نرمال است، آن‌گاه متغیر تصادفی $Z = \frac{p-P}{\sigma(p)}$ توزیع نرمال استاندارد دارد. بدیهی است $E(p) = P$، $\sigma(p) = \sqrt{\frac{N-n}{N-۱} \cdot \frac{PQ}{n}}$. اگر قرار دهیم $\phi_Z(z) = ۱ - \alpha/۲$، آن‌گاه

$$\Pr(-z\sigma(p) < p - P < z\sigma(p)) = ۱ - \alpha$$

یا

$$\Pr(|p - P| \geq z\sigma(p)) = \alpha$$

از مقایسهٔ این رابطه با (الف)، نتیجه می‌شود که باید داشته باشیم $d = z\sigma(p)$، یعنی

$$d = z\sqrt{\frac{N-n}{N-۱} \cdot \frac{PQ}{n}}$$

که اگر آن را برحسب n حل کنیم

$$n = \frac{z^۲PQ/d^۲}{۱ + \frac{۱}{N}\left(\frac{z^۲PQ}{d^۲} - ۱\right)}$$

در کاربردها، باید به‌کمک نمونه‌ای مقدماتی به‌حجم $n_۱$ که متناسب با بودجه موجود تهیه می‌شود P را برآورد کنیم اگر این برآورد، p باشد برآورد Q برابر با $q = ۱ - p$ است. مقدار α از قبل معلوم است و مقدار z نیز با توجه به ضریب اطمینان مفروض $۱ - \alpha$، مشخص می‌شود، لذا

$$\hat{n} = \frac{z^۲pq/d^۲}{۱ + \frac{۱}{N}\left(\frac{z^۲pq}{d^۲} - ۱\right)} \tag{۴۱.۲}$$

اگر قرار دهیم $v = \frac{d^۲}{z^۲}$،

$$\hat{n} = \frac{pq/v}{۱ + \frac{۱}{N}\left(\frac{pq}{v} - ۱\right)}$$

اگر N بزرگ باشد $\hat{n}$ تقریباً برابر می‌شود با

$$\hat{n}_۰ = \frac{pq}{v}$$

از برابری بالا رابطهٔ تقریبی $v = \frac{pq}{\hat{n}_۰}$ نتیجه می‌شود که v به‌همان صورت واریانس تقریبی نسبت نمونه‌ای است. در عمل ابتدا $\hat{n}_۰$ را حساب می‌کنند. اگر $\hat{n}_۰/N$ قابل اغماض باشد $\hat{n}_۰$ همان برآورد رضایتبخش n است و اگر قابل اغماض نباشد، از رابطهٔ

$$\hat{n} = \frac{\hat{n}_۰}{۱ + \frac{\hat{n}_۰ - ۱}{N}} \simeq \frac{\hat{n}_۰}{۱ + \frac{\hat{n}_۰}{N}} \tag{۴۲.۲}$$

استفاده می‌کنند.

مثال ۲۱.۲ می‌خواهند در ناحیه‌ای نسبت افرادی را که گروه خونی A دارند برآورد کنند. از سابقهٔ موجود در درمانگاه ناحیه، نسبت $p = ۰٫۵$ به‌دست آمده است. می‌خواهند با یک نمونه‌گیری

تصادفی بدون جایگذاری برآوردی برای نسبت دارندگان گروه خونی A در ناحیه، یعنی برآوردی برای P به‌دست آورند به‌قسمی

$$\Pr(|\hat{P} - P| \geq ۰٫۰۵) = ۰٫۰۵$$

در این صورت حجم نمونه را چقدر باید برآورد کنند؟

در این مثال $d = ۰٫۰۵$، $\alpha = ۰٫۰۵$، $z \simeq ۲$، بنابراین

$$v = \frac{d^۲}{z^۲} = \frac{(۰٫۰۵)^۲}{۴} = \frac{۱}{۱۶۰۰}$$

لذا اگر تعداد ساکنین ناحیه را بزرگ فرض کنیم حجم نمونه به‌صورت زیر برآورد می‌شود

$$\hat{n}_۰ = \frac{pq}{v}$$

p نسبتی است که از روی نمونه‌گیری مقدماتی به‌دست می‌آید. در این مثال سابقهٔ موجود در درمانگاه را به‌عنوان نمونهٔ مقدماتی می‌پذیریم و لذا $p = ۰٫۵$، $q = ۰٫۵$، پس

$$\hat{n}_۰ = \frac{(۰٫۵)(۰٫۵)}{۱/۱۶۰۰} = ۴۰۰$$

اگر تعداد ساکنین ناحیه مثلاً ۳۲۰۰ باشد، حجم نمونه به‌صورت زیر برآورد می‌شود

$$\hat{n} = \frac{\hat{n}_۰}{۱ + (\hat{n}_۰ - ۱)/N} = \frac{۴۰۰}{۱ + ۳۹۹/۳۲۰۰} = ۳۵۶$$

▲

چند تذکر. ۱. گاهی اوقات، به‌خصوص وقتی می‌خواهیم NP، تعداد کل واحدهایی از جامعه را که در ردهٔ C قرار دارند، برآورد کنیم مایلیم به‌جای خطای مطلق، خطای نسبی را کنترل نماییم. می‌دانیم برآورد NP برابر با Np است. خطای نسبی برابر با $\left|\frac{Np-NP}{NP}\right|$ است. می‌خواهیم حجم نمونه را برآورد کنیم به‌شرطی که

$$\Pr\left(\left|\frac{Np - NP}{NP}\right| \geq r\right) = \Pr(|p - P| \geq rP) = \alpha$$

در این صورت کافی است برای برآورد n در (۴۱.۲) به‌جای d مقدار rP را قرار دهیم. بدیهی است چون P مجهول است، با توجه به نتایج نمونه‌ای مقدماتی به‌جای rP برآورد آن rp را قرار می‌دهیم. به‌خصوص وقتی N بزرگ است،

$$\hat{n}_۰ = z^۲/r^۲ \cdot q/p \qquad (۴۳.۲)$$

رابطهٔ (۴۲.۲) بدون تغییر باقی می‌ماند.

۲. گاهی در فرمول (۴۱.۲) یا نظایر آن برآورد p از نمونهٔ مقدماتی در دست نیست. مقداری از p که مربوط به «بدترین حالت» است برابر است با $p = ۰٫۵$، که می‌توان آن را برای تعیین $\hat{n}$ به‌کار برد. مقدار pq و در نتیجه $\hat{n}$ حاصل از فرمول، ماکسیمم مقدار خود را به‌ازای $p = ۰٫۵$ اختیار می‌کند.

مثال ۲۲.۲ در دانشکده‌ای به‌کمک نمونه‌ای مقدماتی نسبت دانشجویان متأهل برابر با ۰٫۲۱ برآورد شده است. قصد دارند با نمونه‌گیری تصادفی بدون جایگذاری تعداد دانشجویان متأهل را به‌قسمی برآورد کنند که با احتمال ۰٫۹۵ خطای نسبی برآورد کمتر از ۰٫۲ باشد. تعداد کل دانشجویان برابر با ۱۵۰۰ است. حجم نمونه را برآورد کنید.

در این مثال $r = ۰٫۲$، $p = ۰٫۲۱$، $\alpha = ۰٫۰۵$، $z \simeq ۲$ لذا بدون توجه به تعداد دانشجویان

$$\hat{n}_۰ = \frac{۴}{(۰٫۲)^۲} \frac{۰٫۷۹}{۰٫۲۱} \simeq ۳۷۷$$

▲

تذکر. **برآوردهای $S^۲$ی جامعه در محاسبهٔ حجم نمونه.** برای استفاده از فرمولهای مربوط به محاسبهٔ n، باید $S^۲$ی جامعه معلوم باشد. چون اصولاً $S^۲$ مجهول است همان‌طورکه متذکر شدیم باید به‌کمک نمونه‌ای مقدماتی برآوردی برای آن به‌دست آورد و با قرار دادن این برآورد در فرمول مزبور، برآوردی برای n محاسبه کرد. معمولاً در عمل به‌جای استفاده از نمونهٔ مقدماتی، نمونهٔ اصلی را در دو مرحله انتخاب می‌کنند. نمونهٔ مرحلهٔ اول نمونه‌ای تصادفی به‌حجم $n_۱$ است که از روی آن $s_۱^۲$ را محاسبه می‌کنند و با توجه به این مقدار، و محاسبهٔ $\hat{n}$ و با در نظر گرفتن مطالبی که ذیلاً در چند حالت مطرح می‌شوند مقدار n را برآورد کرده، و $\hat{n} - n_۱$ واحد لازم را به‌تصادف انتخاب می‌نمایند. در تمام حالتها فرض بر این است که $n_۱ \leq n$.

۱. اگر ضریب تغییرات، یعنی نسبت $\frac{S}{\overline{Y}_N}$ یا $\frac{\sigma_Y}{\overline{Y}_N}$ معلوم باشد، قرار می‌دهیم $D = \frac{S^۲}{\overline{Y}_N^۲}$. اگر $s_۱^۲$ از روی نمونه به‌حجم $n_۱$ در مرحلهٔ اول حاصل شود و پذیرهٔ نرمال بودن توزیع جامعه معتبر باشد، ثابت شده است که می‌توان n را از فرمول زیر برآورد کرد [۳].

$$\hat{n} = \frac{s_۱^۲}{D\overline{Y}_۱^۲}\left(۱ + ۸D + \frac{s_۱^۲}{n_۱\overline{Y}_۱^۲} + \frac{۲}{n_۱}\right)$$

که در آن $\overline{Y}_۱$ میانگین نمونهٔ مرحلهٔ اول است. $\overline{Y}_n$ که میانگین نمونهٔ به‌حجم $\hat{n}$ است نااریبی مختصری دارد. سرانجام میانگین جامعه را از فرمول $\hat{\overline{Y}}_N = \overline{Y}_n(۱ - ۲D)$ برآورد می‌کنند.

۲. اگر بخواهیم $\hat{\overline{Y}}_N$ را به‌گونه‌ای بیابیم که واریانس آن برابر Vی معلوم باشد ثابت می‌شود که

اگر در مرحلهٔ اول، از روی نمونهٔ به‌حجم $n_۱$، مقدار $s_۱^۲$ حاصل شود، آن‌گاه

$$\hat{n} = \frac{s_۱^۲}{\mathrm{V}}\left(۱ + \frac{۲}{n_۱}\right)$$

اگر S دقیقاً معلوم باشد حجم نمونهٔ مورد نیاز برابر با $\frac{S^۲}{\mathrm{V}}$ است [۳].

۳. اگر بخواهیم $\hat{n}$ را وقتی نسبت P در جامعه برآورد می‌شود به‌گونه‌ای بیابیم که V، واریانس $\hat{P}$ معلوم باشد، ثابت که می‌شود که وقتی $p_۱$ برآورد P از نمونهٔ مرحلهٔ اول به‌حجم $n_۱$ است آن‌گاه

$$\hat{n} = \frac{p_۱q_۱}{\mathrm{V}} + \frac{۳ - ۸p_۱q_۱}{p_۱q_۱} + \frac{۱ - ۳p_۱q_۱}{\mathrm{V}n_۱}$$

۱۹.۲ نمونه‌گیری وارون

اگر نسبت p در جامعه خیلی کوچک باشد (که غالباً چنین است)، صفت تحت بررسی طبیعتاً صفتی کمیاب خواهد بود. در این موارد روش برآورد p به‌طریق معمول چندان رضایتبخش نیست. مثلاً تصور کنید که بخواهیم نسبت مردان بالغی را که در یک ناحیه مبتلا به هموفیلی هستند، بدانیم. اگر تصمیم بگیریم از N واحد جامعه، نمونه‌ای تصادفی به‌حجم n انتخاب کنیم و به‌کمک آن نسبت مورد نظر را برآورد نماییم، چون بیماری هموفیلی بیماری شایعی نیست چه‌بسا که در نمونهٔ n واحدی اصلاً فردی مبتلا نیابیم و نتیجه بگیریم که برآورد نسبت بیماران صفر است، در حالی که احتمال دارد از نمونهٔ n واحدی دیگری چنین برآوردی به‌دست نیاید. در چنین مواردی می‌توان از روشی که به نمونه‌گیری وارون موسوم است استفاده کرد. در این روش که منسوب به فینی[1] است، حجم نمونه را از قبل تثبیت نمی‌کنند. برعکس از قبل تصمیم می‌گیرند که نمونه‌گیری را تا ظاهر شدن m واحد دارای صفت کمیاب ادامه دهند. وقتی m واحد ظاهر شدند تعداد واحدهایی که تا این زمان برای یافتن واحدهای با صفت کمیاب بررسی کرده‌اند برابر با n بوده و معرف حجم نمونه است که بالقوه یک متغیر تصادفی است. اگر P نسبت واقعی صفت در جامعهٔ به‌حجم N باشد تعداد واحدهای واجد صفت کمیاب در جامعه برابر NP است. هدف، تعیین برآورد P و برآورد NP است. ثابت می‌کنند که برآورد نااریب P به‌صورت زیر است

$$\hat{P} = p = \frac{m - ۱}{n - ۱} \qquad n \geq m \tag{۴۳.۲}$$

برآورد نااریب واریانس برآوردکننده بالا نیز به‌صورت زیر به‌دست می‌آید

$$\hat{\mathrm{V}}(\hat{P}) = \frac{m - ۱}{n - ۱}\left[\frac{m - ۱}{n - ۱} - \frac{N - ۱}{N}\cdot\frac{m - ۲}{n - ۲} - \frac{۱}{N}\right] \tag{۴۴.۲}$$

1. Finney

اگر N بزرگ باشد می‌توان ثابت کرد که

$$\hat{P} = p = \frac{m-1}{n-1}$$

$$\hat{\mathrm{V}}(\hat{P}) = \frac{p(1-p)}{n-2}$$

قبل از اینکه این فصل را به پایان ببریم دربارهٔ نمونه‌گیری صید و بازصید که در ارتباط با نمونه‌گیری تصادفی ساده است به خواننده ایده‌ای می‌دهیم.

۲۰.۲ نمونه‌گیری صید و بازصید

نمونه‌گیری صید و بازصید برای برآورد کردن تعداد کل افراد یک جامعه به‌کار می‌رود. در این نمونه‌گیری ابتدا نمونه‌ای از جامعه به‌تصادف انتخاب می‌کنند، و سپس واحدهای این نمونه را علامت‌گذاری کرده به جامعه برمی‌گردانند. در مرحلهٔ دوم نمونه‌ای به‌تصادف از جامعه استخراج می‌کنند. این نمونه مستقل از نمونهٔ اول است. در این نمونهٔ دوم، تعداد واحدهایی را که علامت گذاری شده‌اند مشخص می‌نمایند. اگر نمونهٔ دوم نمونهٔ نماینده‌ای از کل جامعه باشد نسبت نمونه‌ای واحدهای علامت‌گذاری‌شده در این نمونه برآوردی از نسبت واحدهای علامت‌گذاری‌شده در کل جامعه است. از این بستگی می‌توان استفاده و تعداد کل واحدهای جامعه را برآورد کرد. در این نمونه‌گیری، نمونهٔ اول را صید و نمونه دوم را که شامل برخی از واحدهای نمونهٔ اول است بازصید می‌گویند.

نمونه‌گیری صید و بازصید برای برآورد فراوانی جامعه‌های حیوانات، پرندگان، ماهیها، خزندگان، حشرات و غیره و همچنین برای برآورد پارامترهای بقا به‌کار می‌رود. گاهی در جامعه‌های انسانی برای برآورد تعداد بی‌مسکنها، بیکارها و تعداد پیشامدهایی نظیر تعداد تصادفها از این نمونه‌گیری استفاده می‌شود. در نمونه‌گیری از جامعهٔ حیوانات معمولاً واحدهای نمونهٔ اول را به‌طریقی صید می‌کنند و آنها را علامت‌گذاری یا به دستگاه فرستنده مجهز می‌کنند. در جامعه‌های انسانی دو نمونه درواقع دو فهرست از اسمها هستند. مثلاً فهرست اول، داده‌هایی است که از سرشماری انتخاب شده‌اند و فهرست دوم داده‌هایی از نمونه‌گیری مقطعی بعد از سرشماری است. یا فهرست اول، فهرست ثبت حوادث بخش سوانح یک بیمارستان و فهرست دوم ثبتهای مربوط به شرکت بیمه است. در زمینهٔ این نمونه‌گیری، کورماک[1] (۱۹۷۹)، وایت و اندرسن[2] (۱۹۷۸)، پولوک[3] (۱۹۸۱، ۱۹۹۱)، برونی[4] (۱۹۹۰) مراجع مفیدی هستند. اخیراً روشهایی برای به‌کار گرفتن صید و بازصید در جامعه‌های انسانی در مقاله‌هایی از کوان[5] و مالک (۱۹۸۶)، فریدمن (۱۹۹۱)، سودمن[6]، سیرکن[7] و کوان (۱۹۸۸) و ولتر[8] (۱۹۹۱) عرضه شده‌اند.

در خلاصه‌ای که برای روش نمونه‌گیری صید و بازصید می‌آوریم تعداد کل افراد جامعه را که مجهول است و باید برآورد شود با τ نشان می‌دهیم. تعداد کل واحدهای اولین نمونه را که همهٔ آنها

1. Cormack 2. White and Anderson 3. Pollock 4. Brownie 5. Cowan
6. Sudman 7. Sirken 8. Wolter

علامت‌گذاری می‌شوند با نماد X معرفی می‌کنیم. تعداد افراد نمونهٔ دوم را با نماد y نشان می‌دهیم و تعداد واحدهایی از نمونهٔ دوم راکه علامت دارند و متغیری تصادفی است با x نمایش می‌دهیم.

۲.۲۰.۱ برآوردکننده‌های حجم جامعه

در بررسی صید-بازصید از جامعهٔ حیوانات، یک نمونهٔ متشکل از X حیوان دستگیر و پس از علامت‌گذاری در جامعه رها می‌شوند. سپس از کل جامعه y حیوان شکارشده و پس از بررسی، x، تعداد حیواناتی که علامت‌دار هستند معین می‌شود. اگر نمونهٔ دوم نمایندهٔ جامعه باشد نسبت حیوانات علامت‌دار در نمونه، یعنی $\frac{x}{y}$ تقریباً برابر با نسبت حیوانات علامت‌دار درکل جامعه است. اما تعداد حیوانات علامت‌دار درکل جامعه برابر با X، یعنی حجم نمونهٔ اول است و نسبت حیوانات علامت‌دار درکل جامعه $\frac{X}{\tau}$ است (τ تعداد واحدهای جامعه و مجهول است). پس

$$\frac{x}{y}=\frac{X}{\tau}$$

این برابری، یک برابری تقریبی است که از حل آن نسبت به τ داریم

$$\hat{\tau}=\frac{y}{x}X$$

این برآوردکننده به برآوردکنندهٔ پترسن[1] معروف است. سکار[2] و دمینگ[3] برآوردکنندهٔ واریانس $\hat{\tau}$ را به‌صورت زیر ارائه داده‌اند

$$\hat{\mathrm{V}}(\hat{\tau})=\frac{Xy(X-x)(y-x)}{x^{۳}}$$

یک بازهٔ اطمینان $(۱-\alpha)$ ۱۰۰ درصد برای τ به‌صورت زیر است

$$\hat{\tau}\pm z\sqrt{\hat{\mathrm{V}}(\hat{\tau})}$$

که در آن $\phi(z)=۱-\frac{\alpha}{۲}$ و ϕ تابع توزیع نرمال استاندارد است. چون ممکن است در نمونهٔ دوم هیچ واحد علامت‌داری وجود نداشته باشد لذا x می‌تواند مقدار صفر را هم اختیار کند و لذا $\hat{\tau}$ نامتناهی و واریانس $\hat{\tau}$ هم نامتناهی است. با اندیشهٔ رفع این مشکل، چپمن[4] برآوردکنندهٔ اصلاح‌شده زیر را پیشنهاد کرده است (اثر سبر، ۱۹۸۲ را ببینید).

$$\tilde{\tau}=\frac{(X+۱)(y+۱)}{(x+۱)}-۱$$

1. Petersen 2. Sekar 3. Deming 4. Chapman

برآوردکنندهٔ تقریباً نااریب واریانس برای برآوردکنندهٔ چپمن به‌صورت زیر است

$$\hat{V}(\tilde{\tau})=\frac{(X+1)(y+1)(X-x)(y-x)}{(x+1)^2(x+2)}$$

یک بازهٔ اطمنیان $100(1-\alpha)$ درصد برای τ با استفاده از برآوردکنندهٔ چپمن به‌صورت

$$\tilde{\tau}\pm z\sqrt{\hat{V}(\tilde{\tau})}$$

است، که در آن $\phi(z)=1-\frac{\alpha}{2}$ و ϕ تابع توزیع نرمال استاندارد است. شیوه‌های دیگر تعیین بازهٔ اطمینان برای دو برآوردکنندهٔ τ، خصوصاً تحت پذیرهٔ توزیعی مفروض وجود دارند. ویژگیهای بزرگ نمونه‌ای برآوردکننده‌های نمونه‌گیری صید-بازصید را می‌توان در اثر سن[1] (۱۹۸۸) مرور کرد.

مثال ۲۳.۲ در ناحیه‌ای $X=200$ خرگوش را به دام انداخته و پس از علامت‌گذاری رها کرده‌ایم. چند روز بعد به ناحیهٔ مذکور مراجعت کرده و مستقل از دام‌گذاری قبل، $y=120$ خرگوش به دام انداخته و دریافتیم که $x=40$تای آنها علامت دارند.

برآورد تعداد خرگوشها به‌روش پترسن در جامعهٔ مربوط به این ناحیه به‌صورت زیر حساب می‌شود

$$\hat{\tau}=\frac{120}{40}(200)=600$$

برآورد واریانس برآوردکنندهٔ $\hat{\tau}$ به‌صورت زیر محاسبه می‌شود

$$\hat{V}(\hat{\tau})=\frac{200(120)(200-40)(120-40)}{40^3}=4800$$
$$\hat{\sigma}(\hat{\tau})=68{,}4$$

و بازهٔ اطمینان ۹۵ درصد برای τ به‌صورت زیر است

$$600\pm 1{,}96(68{,}4):\qquad (465\ ،\ 735)$$

اگر از برآوردکنندهٔ چپمن استفاده کنیم

$$\tilde{\tau}=\frac{(200+1)(120+1)}{(40+1)}-1\#593$$

برآورد واریانس $\tilde{\tau}$ به‌صورت زیر حساب می‌شود

$$\hat{V}(\tilde{\tau})=\frac{(200+1)(120+1)(200-40)(120-40)}{(40+1)^2(40+2)}\#4409{,}35$$

1. Sen

و

$$\hat{\sigma}(\hat{\tau}) = ۶۶٫۴$$

که می‌توان به‌کمک این برآوردکننده نیز بازهٔ اطمینان ۹۵ درصد برای τ یافت. این بازه به‌صورت زیر است

$$۵۹۳ \pm ۱٫۹۶(۶۶٫۴) : (۴۶۲ \ , \ ۷۲۴)$$

۵

۲.۲۰.۲ مدلهای صید‌ـ‌بازصید ساده

برای صید‌ـ‌بازصید از جامعه‌ای بسته تحت مدل چندجمله‌ای، هر حیوان جامعه را می‌توان دقیقاً در چهار رسته، رسته‌بندی کرد. صیدشده در هر دو نمونه، صیدشده در نمونهٔ اول و نه در نمونهٔ دوم، صیدشده در نمونهٔ دوم و نه در نمونهٔ اول، و بالاخره صیدنشده در هیچ‌یک از دو نمونه. وقوع هر یک از این چهار برآمد احتمال خاص خود را دارد. لذا مدل چندجمله‌ای با چهار احتمال مربوط به چهار رسته، مدلی است که برای سابقه هر حیوان جامعه مدلی مناسب است. فرض می‌کنیم در طول هر نمونه‌گیری شانس صید هر فرد با فرد دیگر یکسان اما در دو نمونه‌گیری احتمال صید متفاوت باشد. ضمناً می‌پذیریم که دو نمونه مستقل از یکدیگرند. با چنین مدلی، برآوردکننده درستنمایی ماکسیمم مجموع جامعه برابر با جزء صحیح برآوردکنندهٔ پترسن، یعنی $\hat{\tau} = Xy/x$ است. برآوردکنندهٔ ML احتمال $p_۱$، یعنی احتمال صید در نمونهٔ اول برابر با $\hat{p}_۱ = \frac{x}{y}$. برآوردکنندهٔ ML احتمال $p_۲$ یعنی احتمال صید در نمونهٔ دوم برابر با $\hat{p}_۲ = \frac{x}{X}$ است. آهلو[1] (۱۹۹۰) نشان داده است که حتی اگر احتمالهای صید برای افراد مختلف در نمونهٔ اول متفاوت اما برابر با احتمال صید در نمونهٔ دوم باشند هنوز برآوردکنندهٔ پترسن صادق است.

اگر احتمال یک تک صید برای هر فرد در هر دو نمونه برابر p باشد و نمونه‌ها از هم مستقل باشند، برآوردکننده‌های ML مقادیر p و τ عبارت‌اند از $\hat{p} = ۲x/(X+y)$ و $\hat{\tau} = (X+y)/۲\hat{p}$.

اگر اعداد X و y در دو نمونه تثبیت‌شده فرض شوند و نمونهٔ دوم، نمونهٔ تصادفی ساده بدون جایگذاری از افراد جامعه باشد، آن‌گاه عدد x که تعداد حیوانات علامت‌گذاری‌شده در نمونهٔ دوم است دارای توزیع ابر هندسی است. با فرض برابری احتمال صید همهٔ افراد، این توزیع، توزیع شرطی x تحت مدل چندجمله‌ای به‌شرط X و y است. با این مدل، برآوردکنندهٔ ML مقدار τ، باز هم جزء صحیح برآوردکنندهٔ پترسن است. وقتی نمونه‌گیری وارون مورد نظر باشد باید x را از قبل تثبیت شده گرفت. در این صورت y متغیری تصادفی است و نمونه‌گیری تا صید x فرد علامت‌دار ادامه می‌یابد.

تمرینها

۱ . تعداد فرزندان ۶ کارمند یک بخش به‌ترتیب ۱، ۳، ۲، ۰، ۴، ۵ است.

1. Ahlo

الف) به‌عنوان تمرین همهٔ نمونه‌های تصادفی بدون جایگذاری و با حجم ۴ از این جامعه را بنویسید. میانگینهای آنها را به‌دست آورید و نشان دهید که متوسط این میانگینها با میانگین جامعه برابر است.

ب) واریانس میانگینهای نمونه‌ای را حساب کنید و درستی رابطهٔ

$$\mathrm{V}(\overline{Y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

را تحقیق کنید.

ج) همه s^2های نمونه‌ای را حساب کنید و درستی برابر $E(s^2) = S^2$ را تحقیق کنید.

د) فرض کنید که ۶ واحد جامعه را نداریم و تنها نمونهٔ تصادفی بدون جایگذاری ۴، ۰، ۲، ۳ را در دست داریم. از روی این نمونه برآوردی نااریب برای میانگین جامعه و برآوردی نااریب برای واریانس جامعه ارائه دهید. برای $\mathrm{V}(\overline{Y}_2)$ نیز برآوردی نااریب به‌دست آورید.

ه) با توجه به قسمت (د)، برآوردی نااریب برای مجموع واحدهای جامعه به‌دست آورید، و برآوردی نااریب برای واریانس این برآوردکننده معرفی کنید.

۲. از جامعه چهار واحدی با مقادیر ۲، ۳، ۴، ۱ همهٔ نمونه‌های به‌حجم ۲ و با جایگذاری را تعیین کنید.

الف) مستقیماً تحقیق کنید که متوسط میانگینهای نمونه‌ها برابر با میانگین جامعه است.

ب) s^2ی همهٔ نمونه‌ها را معین کنید و درستی برابری $\sigma^2 = E(s^2)$ را تحقیق کنید. σ^2 واریانس جامعه است.

ج) واریانس میانگینهای نمونه‌ای را حساب کرده، درستی برابری $\mathrm{V}(\overline{Y}_n) = \sigma^2/n$ را تحقیق کنید.

د) فرض کنید همهٔ واحدهای جامعه را نداریم و تنها نمونهٔ با جایگذاری ۲، ۳ را داریم. از روی این نمونه برآوردی نااریب برای میانگین جامعه به‌دست آورید، و برآورد نااریب واریانس $\overline{Y}_n$ را حساب کنید.

۳. به‌کمک جدول اعداد تصادفی از جامعه‌ای که واحدهای آن از ۱ تا ۷۷۵ شماره‌گذاری شده‌اند، نمونه‌ای به‌حجم ۱۵ انتخاب کنید.

۴. دو دندانپزشک A و B وضعیت دندانهای ۲۰۰ کودک زیر ۷ سال دهکده‌ای را بررسی می‌کنند. دندانپزشک A، نمونه‌ای تصادفی بدون جایگذاری شامل ۲۰ کودک انتخاب می‌کند و تعداد دندانهای خراب هر کودک را مشخص و شمارش می‌کند. نتیجهٔ کار او در جدول زیر آمده است:

تعداد دندانهای خراب هر کودک :	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
تعداد کودکان :	۸	۴	۲	۲	۱	۱	۰	۰	۰	۱	۱

دندانپزشک B با همان تکنیک معاینه، همهٔ ۲۰۰ کودک را معاینه و تنها آنهایی را متمایز می‌کند که دندانهای خراب ندارند و درمی‌یابد که جمعاً ۶۰ کودک بدون دندان خراب‌اند. تعداد دندانهای خراب کودکان دهکده را

الف) با استفاده از نتایج معاینات دندانپزشک A،

ب) با استفاده از نتایج کار هر دو دندانپزشک،

برآورد کنید. آیا این دو برآورد نااریب‌اند؟ دقت کدام برآورد بیشتر است؟

۵ . در دانشکده‌ای با ۲۰۰۰ دانشجو، می‌خواهند نظر دانشجویان را دربارهٔ تغییر سیستم نیمسالی به ثلثی بدانند. ۴۰ دانشجو به‌تصادف انتخاب می‌کنند و نتیجه می‌گیرند که ۲۵ نفر موافق تغییرند. نسبت مخالفان تغییر را برای کل دانشجویان برآورد کنید و برای این برآوردکننده بازهٔ اطمینانی با ضریب اطمینان ۹۵ر۰ بیابید.

۶ . تعداد مراجعان به یک آزمایشگاه در طول ماه ۱۷۵۰ نفر بوده است. ضمن آزمایشهای دیگر، میزان قند خون همهٔ مراجعان اندازه‌گیری شده است. نمونه‌ای به‌حجم ۱۷۵ از نتایج آزمایش این بیماران انتخاب کرده‌اند و نتیجه گرفته‌اند که ۱۵ نفر آنها به دیابت مبتلا هستند. تعداد بیماران دیابتی را در کل جامعهٔ مراجعان ماه برآورد کنید و بازهٔ اطمینانی با ضریب اطمینان ۹۰ر۰ برای این برآوردکننده به‌دست آورید.

۷ . فرض کنید ۵ دانشجوی A، B، C، D، و E داریم. اگر تهرانی بودن را با ۱ و شهرستانی بودن را با ۰ نشان دهیم، داریم

نام دانشجو :	A	B	C	D	E
تهرانی یا شهرستانی :	۰	۱	۱	۰	۱

اگر این دانشجویان یک جامعه به‌حساب آیند نسبت تهرانیها را تعیین کنید و سپس در جدولی همهٔ نمونه‌های دوتایی بدون جایگذاری ممکن را مشخص کنید و نسبت تهرانیها را برای هر نمونه (p) به‌دست آورید و به‌کمک آنها درستی نااریبی برآوردکنندهٔ p را تحقیق کنید. از روی توزیع p، واریانس p را معین و درستی برابری $\mathrm{V}(p) = \frac{N-n}{N-1}\frac{PQ}{n}$ را تحقیق کنید. درستی برابری $E(\hat{\mathrm{V}}(p)) = \mathrm{V}(p)$ را بررسی نمایید.

۸ . امضاهای مربوط به یک درخواست در ۶۷۶ برگ جمع‌آوری شده‌اند. هر برگ فضای لازم برای ۴۲ امضا دارد ولی در بعضی از برگها تعداد امضاها کمتر است. ۵۰ برگ به‌تصادف انتخاب شده و نتیجه شمارش امضاهای آنها به‌صورت جدول زیر بوده است

تعداد امضا، Y_i :	۴۲	۴۱	۳۶	۳۲	۲۹	۲۷	۲۳	۱۹	۱۶
فراوانی، f_i :	۲۳	۴	۱	۱	۱	۲	۱	۱	۲

تعداد امضا، Y_i :	۱۵	۱۴	۱۱	۱۰	۹	۷	۶	۵	۴	۳
فراوانی، f_i :	۲	۱	۱	۱	۱	۱	۳	۲	۱	۱

مجموع کل امضاها را برآورد کنید و بازهٔ اطمینانی با ضریب اطمینان ۹ر۰ برای مجموع امضاها به‌دست آورید.

۹ . نه دهکدهٔ به‌ترتیب دارای ۶۹۳، ۱۶۰، ۸۷۰، ۵۵۷، ۱۶۲۱، ۱۵۰۳، ۷۶۴، ۸۸۳، و ۸۱۶ مزرعه هستند. می‌خواهیم از مزارع دهکده‌ها ۶ مزرعه به‌عنوان نمونه انتخاب کنیم. به‌کمک جدول

اعداد تصادفی این نمونه را مشخص کنید.

۱۰ . مدیر یک مرغداری می‌خواهد وزن کل $N = ۱۰۰۰$ جوجهٔ ۴ هفته‌ای را برآورد کند. چه تعداد از جوجه‌ها را باید به‌تصادف انتخاب کند تا بازهٔ اطمینان با ضریب ۹۵ر۰ به‌طول ۲۰۰ کیلوگرم باشد. مطالعه‌های مشابهی که با شرایط مشابه درگذشته انجام شده‌اند حاکی از این است که $S^۲$ی جامعهٔ وزنها تقریباً $\frac{۱}{۲}$ کیلوگرم است. کسر نمونه‌گیری قابل اغماض است.

۱۱ . نمونه‌ای تصادفی به‌حجم ۵۰ از ۱۰۰۰ دبیر درس ریاضیات گسسته انتخاب کرده‌اند و نظر آنها را دربارهٔ لزوم این درس در دورهٔ پیشدانشگاهی خواسته‌اند. ۲۲ نفر موافق، ۱۹ نفر مخالف بوده و بقیه عقیدهٔ خاصی نداشته‌اند. تعداد موافقان را در کل جامعهٔ دبیران این درس برآورد کنید. واریانس این برآوردکننده را برآورد کنید و برای تعداد موافقان در جامعه، بازهٔ اطمینانی با ضریب ۸ر۰ به‌دست آورید.

۱۲ . از جامعه‌ای به‌حجم N، دو نمونهٔ تصادفی بدون جایگذاری، مستقل از هم و به‌ترتیب به حجمهای $n_۱$ و $n_۲$ $(n_۱ > n_۲)$ استخراج می‌کنیم. میانگینهای دو نمونه را با $\overline{Y}_۱$ و $\overline{Y}_۲$ نشان می‌دهیم. برای برآورد میانگین جامعه، آمارهٔ $\frac{۱}{۲}(\overline{Y}_۱ + \overline{Y}_۲)$ را به‌کار می‌بریم. تحت چه شرطی این آماره دقیقتر از $\overline{Y}_۱$ و $\overline{Y}_۲$ میانگین جامعه را برآورد می‌کند؟

۱۳ . در شهرکی ۱۲۰۰ خانوار زندگی می‌کنند. می‌خواهیم با تهیهٔ نمونه‌ای تصادفی، میانگین درآمد ماهیانهٔ آنها را برآورد کنیم. مایلیم با احتمال ۹۵ر۰، قدر مطلق تفاضل برآورد و میانگین واقعی از ۰۴ر۰ میانگین کوچکتر باشد. حجم نمونه را چقدر باید انتخاب کنیم؟ در یک بررسی مقدماتی میانگین و S جامعه برحسب ۱۰۰۰ تومان برابر با ۱۶ و ۴ برآورد شده‌اند.

۱۴ . هزار لامپ خریداری شده است. برای برآورد زیان حاصل از خرید لامپهای معیوبی که بین لامپها وجود دارند ۵۰ تا از آنها را به‌تصادف انتخاب کرده و پس از آزمایش دریافتیم که ۱۲ تا از آنها معیوب‌اند. اگر بهای هر لامپ ۲۴۰ ریال باشد برآورد کل زیان حاصل از خرید را به‌دست آورید.

۱۵ . از جامعه‌ای به‌حجم N نمونه‌ای تصادفی به‌حجم n انتخاب می‌کنیم، سپس از این نمونه، زیرنمونه‌ای تصادفی به‌حجم $n_۱ < n$ می‌گیریم. میانگین نمونه را $\overline{Y}$ و میانگین زیرنمونه را $\overline{Y}_۱$ می‌نامیم. اگر $\overline{Y}_۲$ میانگین واحدهای باقی‌مانده از نمونهٔ اول پس از استخراج زیرنمونه باشد روابط زیر را ثابت کنید

خیلی مهم

الف) $$\mathrm{V}(\overline{Y}_۱ - \overline{Y}) = \left(\frac{۱}{n_۱} - \frac{۱}{n}\right) S^۲$$

ب) $$\mathrm{V}(\overline{Y}_۱ - \overline{Y}_۲) = \left(\frac{۱}{n_۱} + \frac{۱}{n_۲}\right) S^۲$$

ج) $$\mathrm{cov}(\overline{Y},\ \overline{Y}_۱ - \overline{Y}) = ۰$$

که در آنها، $S^۲$ معرف تغییرات جامعه است.

۱۶ . از جامعه‌ای به‌حجم N نمونه‌ای به‌حجم n به‌تصادف و بدون جایگذاری انتخاب می‌کنیم. اگر مقادیر متغیر تحت بررسی در نمونه $Y_۱$، $Y_۲$، ... ، Y_n بوده و a_i، $i = ۱, \ldots, n$ اعدادی ثابت باشند، آن‌گاه

الف) تحت چه شرطی آمارهٔ $\sum_{i=۱}^{n} a_i Y_i$ برآوردکنندهٔ نااریب جامعه است؟

ب) اگر $\hat{\overline{Y}}_N = \sum_{i=۱}^{n} a_i Y_i$ این برآوردکنندهٔ نااریب باشد ثابت کنید

$$\mathrm{V}(\hat{\overline{Y}}_N) = \frac{S^۲}{N}\left(N + \sum_{i=۱}^{n} a_i^۲ - ۱\right)$$

۱۷ . مسؤول آموزش دانشکده‌ای که ۴۰۰ دانشجو دارد ۱۰ دانشجو را به تصادف انتخاب می‌کند و تعداد نمرات زیر دوازدهٔ ترم گذشته آنها را می‌شمارد. نتیجهٔ شمارش او در جدول زیر خلاصه شده است

تعداد نمرات زیر ۱۲ :	۲	۳	۱۱	۰
تعداد دانشجو :	۱	۲	۲	۵

الف) تعداد نمرات زیر دوازدهٔ همهٔ دانشجویان را برآورد کنید.

ب) اگر مسؤول آموزش بداند که ۵۰ درصد دانشجویان نمره‌ای زیر ۱۲ ندارند با استفاده از این اطلاع و نتیجهٔ شمارش قبلی، مجدداً نمرات زیر دوازدهٔ همهٔ دانشجویان را برآورد کنید.

۱۸ . با یک نمونه‌گیری مقدماتی به‌روش تصادفی ساده از جامعه‌ای به‌حجم ۱۰۰۰، ضریب تغییرات را برابر با $\frac{۱}{۳۰}$ برآورد کرده‌ایم. اگر بخواهیم در نمونه‌گیری اصلی از این جامعه، خطای نسبی با احتمال ۹۷۵ر۰ از یک واحد تجاوز نکند حجم نمونه را چقدر باید انتخاب کنیم؟

۱۹ . از جامعه‌ای به‌حجم N، نمونه‌ای تصادفی به‌حجم n انتخاب می‌کنیم. از این نمونه، زیر نمونه‌ای تصادفی به‌حجم $\frac{n}{۲}$ می‌گیریم. اگر واریانس زیرنمونه را $s'^۲$ بنامیم ثابت کنید $E(s'^۲) = S^۲$، که در آن $S^۲$، تغییرات جامعه است. اگر میانگین نمونه $\overline{Y}$ و میانگین زیرنمونه $\overline{Y}'$ باشد ثابت کنید

$$\mathrm{V}(\overline{Y} - \overline{Y}') = \frac{S^۲}{n}$$

۲۰ . جامعهٔ سه واحدی با مقادیر $Y_۱$، $Y_۲$، $Y_۳$ را در نظر بگیرید. ۳ نمونهٔ تصادفی به‌حجم ۲ از این جامعه را به‌صورت

$$s_۱ = \{Y_۱, Y_۲\} \ , \ s_۲ = \{Y_۱, Y_۳\} \ , \ s_۳ = \{Y_۲, Y_۳\}$$

نشان دهید. برای برآورد میانگین جامعه، برآورد خطی $e(s)$ را به‌صورت زیر در نظر بگیرید:

$$e(s_۱) = \frac{۲}{۳}Y_۱ + \frac{۱}{۲}Y_۲$$

$$e(s_۲) = \frac{۱}{۳}Y_۱ + \frac{۱}{۲}Y_۳$$

$$e(s_۳) = \frac{۱}{۲}Y_۲ + \frac{۱}{۲}Y_۳$$

الف) ثابت کنید به‌ازای همهٔ مقادیر $Y_۱$، $Y_۲$، $Y_۳$، برآوردکنندهٔ $e(s)$ نااریب است.
ب) واریانس $e(s)$ را بیابید.
ج) نشان دهید که مقادیری برای $Y_۱$، $Y_۲$، و $Y_۳$ وجود دارند که به‌ازای آنها واریانس $e(s)$ کمتر از واریانس $\overline{Y}_n$ است.
د) به‌خصوص نشان دهید که اگر داشته باشیم $Y_۱ = ۱$، $Y_۲ = ۲$، و $Y_۳ = ۳$، آن‌گاه

$$\mathrm{V}(e) < \mathrm{V}(\overline{Y}_n)$$

۲۱ . جامعه‌ای متناهی به‌حجم N داریم. برای برآورد میانگین آن، یک نمونهٔ تصادفی ساده با جایگذاری می‌گیریم و آن‌قدر به نمونه‌گیری ادامه می‌دهیم تا در این نمونه $n_۱$ واحد متمایز داشته باشیم. اگر فراوانی واحد rام برابر با f_r باشد بدیهی است $\sum_{r=۱}^{n_۱} f_r = n$، که در آن n حجم نمونه است و مقدار آن متغیری تصادفی است. اگر قرار دهیم:

$$\overline{Y}_{n_۱} = \frac{۱}{n_۱}\sum_{r=۱}^{n_۱} Y_r \quad , \quad \overline{Y}_n = \frac{۱}{n}\sum_{r=۱}^{n_۱} f_r Y_r$$

ثابت کنید

الف) $\overline{Y}_n$ و $\overline{Y}_{n_۱}$ هر دو برآوردکننده نااریب $\overline{Y}_N$ (میانگین جامعه) هستند.
ب) $\mathrm{V}(\overline{Y}_n) = E\left(\frac{۱}{n}\sigma_Y^۲\right)$ که در آن $\sigma_Y^۲$ واریانس جامعه است.

ج)
$$E(n) = N\left(\frac{۱}{N} + \frac{۱}{N-۱} + \cdots + \frac{۱}{N-n_۱+۱}\right)$$

د)
$$E\left(\frac{۱}{n}\right) > \frac{۱}{E(n)} > (N-n_۱)[n_۱(N-۱)]^{-۱}$$

و بالاخره نتیجه بگیرید که $\mathrm{V}(\overline{Y}_n) \geq \mathrm{V}(\overline{Y}_{n_۱})$.

۲۲ . حجم جامعه‌ای برابر با N است. مقدار یک واحد جامعه معلوم و برابر با $Y_۱$ است. نمونه‌ای به‌روش تصادفی ساده بدون جایگذاری به‌حجم n از $N - ۱$ واحد دیگر جامعه می‌گیریم. نشان دهید که برآوردکنندهٔ $Y_۱ + (N-۱)\overline{Y}'_n$ واریانسی کوچکتر از $N\overline{Y}_n$ی دارد که مبتنی بر نمونه‌ای تصادفی، بدون جایگذاری و به‌حجم n از کل جامعه است.

۲۳ . زمین‌شناسی می‌خواهد نسبت طلا را در بخش کوچکی از ناحیه‌ای کوهستانی برآورد کند. مایل است نمونه‌ای تصادفی به‌حجم n را که حاصل از بررسی در n نقطهٔ ناحیه است برای تهیهٔ برآورد به‌کار برد، و وجود یا عدم وجود طلا در سنگهای هر نقطه را مشخص کند. حجم نمونه را چقدر انتخاب کند تا برآوردی به‌دست آورد که با احتمال ۹۵ر۰ به فاصله‌ای دورتر از $d = ۰٫۰۵$ از نسبت واقعی جامعه نباشد؟ (راهنمایی: از تذکر ۲ بخش ۱۸.۲ استفاده کنید.)

۲۴ . یک نمونهٔ تصادفی به‌حجم ۳ با جایگذاری از جامعهٔ به‌حجم N انتخاب می‌کنیم.

الف) نشان دهید احتمال اینکه هر سه واحد نمونه یکسان باشند $P_1 = \frac{1}{N^2}$، احتمال اینکه دو واحد نمونه یکسان باشند $P_2 = \frac{3(N-1)}{N^2}$، و احتمال اینکه ۳ واحد نمونه متمایز باشند، $P_3 = \frac{(N-1)(N-2)}{N^2}$.

ب) اگر $\overline{Y}'_n$ را میانگین واحدهای متفاوت در نمونه بگیریم نشان دهید

$$\mathrm{V}(\overline{Y}'_n) = \frac{(2N-1)(N-1)}{6N^2} S^2 \# \left(1 - \frac{f}{2}\right) \frac{S^2}{3}$$

(اگر نمونه a، b، c باشد $\overline{Y}'_n = \frac{a+b+c}{3}$. اگر نمونه a،a، b باشد، $\overline{Y}'_n = \frac{b+a}{2}$ و اگر نمونه a، a، a باشد $\overline{Y}'_n = \frac{a}{3}$ است.)

برای اثبات رابطه بالا، ابتدا نشان دهید که

$$\mathrm{V}(\overline{Y}'_n) = S^2 \left(\frac{N-1}{N} P_1 + \frac{N-2}{2N} P_2 + \frac{N-3}{3N} P_3\right)$$

ج) نشان دهید که اگر $\overline{Y}_n$ میانگین نمونهٔ تصادفی باشد

$$\mathrm{V}(\overline{Y}'_n) < \mathrm{V}(\overline{Y}_n)$$

که در آن، $\overline{Y}_n$ همان میانگین معمولی نمونهٔ تصادفی با جایگذاری است. در اینجا اثبات برای $n = 3$ خواسته شده است. اثبات کلی مطلب برای $n > 3$ را دس-راج[1] عرضه کرده است.

تمرینهای چهارگزینه‌ای

۱ . نمونه‌ای تصادفی بدون جایگذاری از جامعه‌ای استخراج شده است. میانگین نمونه برابر با ۸، حجم نمونه برابر با ۱۰ و مجموع توان دوم مقادیر واحدها برابر با ۱۰۰۰ است. برآورد ناارىب S^2ی جامعه برابر است با

الف) ۴۰ ب) ۳۶ ج) ۴۱ د) ۳۵

۲ . اگر از جامعه‌ای به‌حجم N نمونه‌ای تصادفی بدون جایگذاری به‌حجم $\frac{N}{4}$ بگیریم و اگر S^2ی جامعه برابر با N^2 باشد، در این صورت واریانس میانگین نمونه برابر است با

الف) $3N$ ب) $\frac{3}{N}$ ج) N د) $\frac{1}{N}$

۳ . تعداد نمونه‌های تصادفی بدون جایگذاری به‌حجم ۴ از جامعه‌ای به‌حجم ۱۰ برابر است با

الف) 4^{10} ب) 10^4 ج) ۲۱۰ د) ۴۰

۴ . از جامعه‌ای نمونه‌ای به‌حجم ۱۰ با جایگذاری استخراج کرده‌ایم، اگر داشته باشیم $\sum(X_i - \overline{X})^2 = 450$، که در آن X_i، $i = 1, 2, \dots, 10$ مقادیر واحدهای نمونه و $\overline{X}$ میانگین آنهاست، آنگاه برآورد ناارىب σ^2ی جامعه برابر است با

1. Des-Raj

الف) ۴۵ ب) ۵۰ ج) $\sqrt{۵۰}$ د) $\sqrt{۴۵}$

۵ . برآورد واریانس میانگین نمونه‌ای تصادفی بدون جایگذاری و به‌حجم ۸ از یک جامعه برابر با ۲ است. اگر حجم جامعه ۴۰ باشد برآورد واریانس جامعه برابر است با

الف) ۲۰ ب) ۱۶ ج) $\frac{۳۹}{۲}$ د) ۲

۶ . صد تراشهٔ کامپیوتر، به‌بهای هر تراشه ۲۰۰ ریال خریده‌ایم. ده تراشه به تصادف انتخاب کرده‌ایم. دو تراشه از آنها معیوب بوده است. برآورد میزان ضرری که در کل متحمل شده‌ایم برابر است با

الف) ۴۰۰ ب) ۴۰۰۰ ج) ۲۰ د) ۲۰۰۰

۷ . از جامعه‌ای بزرگ، نمونه‌ای تصادفی بدون جایگذاری گرفته‌ایم. حجم نمونه ۹ است. میانگین نمونه ۲۵ است. انحراف معیار میانگین برابر با ۳ برآورد شده است. در این صورت بازهٔ اطمینان ۹۵ درصد برای میانگین جامعه عبارت است از

الف) (۱۹, ۳۱) ب) (۲۳, ۲۷) ج) (۲۲, ۲۸) د) (۲۱, ۲۶)، ($z_{۰٫۹۷۵} = ۲$)

۸ . در یک نمونه‌گیری تصادفی بدون جایگذاری از جامعه‌ای به‌حجم ۱۰۰ با انحراف معیار ۳، X_i و X_j دو واحد نمونه‌اند. در این صورت $\operatorname{cov}(X_i,\ X_j)$ برابر است با

الف) $-\frac{۱}{۱۱}$ ب) $\frac{۱}{۱۱}$ ج) ۰٫۰۹ د) $\frac{۱}{۲}$

۹ . در جامعه‌ای مقدار انحراف معیار برابر با ۳ و ضریب تغییرات برابر با ۰٫۶ است. در این صورت میانگین جامعه برابر است با

الف) ۱٫۸ ب) ۵ ج) ۲٫۴ د) ۵٫۴

۱۰ . در یک جامعه، یک بار نمونه‌ای به‌حجم ۱۰ به‌روش تصادفی سادهٔ بدون جایگذاری و یک بار نمونه‌ای با همان حجم ۱۰ ولی به‌روش تصادفی سادهٔ با جایگذاری اختیار کرده‌ایم. اگر واریانس برآوردکنندهٔ میانگین جامعه در این دو نمونه‌گیری به‌ترتیب $V_۱$ و $V_۲$ باشد، آن‌گاه

الف) $V_۱ = V_۲$ ب) $V_۱ > V_۲$ ج) $V_۱ \leq V_۲$ د) $V_۱ = \frac{V_۲}{۱۰}$

۱۱ . می‌خواهیم در شهری تعداد بیماران هموفیلی را برآورد کنیم. اگر نسبت این بیماران در شهر برابر P باشد، با اجرای نمونه‌گیری تصادفی ساده بدون جایگذاری آن‌قدر به انتخاب واحدهای نمونه ادامه می‌دهیم تا m (معلوم) فرد هموفیلی مشخص شوند. اگر تا ظهور m فرد هموفیلی، n فرد مورد آزمایش قرار گرفته باشند، آن‌گاه برآورد نااریب P برابر است با

الف) $\hat{P} = \frac{m}{m-۱}$ ب) $\hat{P} = \frac{m}{n}$ ج) $\hat{P} = \frac{m-۱}{n-۱}$ د) $\hat{P} = \frac{m-۱}{n}$

۱۲ . از جامعه‌ای به‌حجم ۱۰۰ نمونه‌ای تصادفی بدون جایگذاری به‌حجم ۲۰ می‌گیریم. مجدداً از این نمونه، نمونه‌ای تصادفی بدون جایگذاری به‌حجم ۱۲ می‌گیریم. میانگینهای این دو نمونه را به‌ترتیب $\overline{Y}$ و $\overline{Y}_۱$ می‌نامیم. در این صورت $\operatorname{cov}(\overline{Y},\ \overline{Y}_۱ - \overline{Y})$ برابر است با

الف) صفر ب) ۱ ج) $\frac{۳}{۵}$ د) $-\frac{۱}{۱۹}$

۱۳ . همهٔ نمونه‌های ممکن به‌حجم n از جامعهٔ به‌حجم N را که به‌روش تصادفی بدون جایگذاری انتخاب می‌شوند، در نظر می‌گیریم. واحدهای جامعه از هم متمایزند. اگر همه واحدهای این نمونه‌ها

را بررسی کنیم هر واحد در همهٔ نمونه‌ها جمعاً به تعداد زیر ظاهر می‌شود

الف) $\binom{N}{n-1}$ ب) $\binom{N-1}{n-1}$ ج) $\binom{N-1}{n}$ د) $\binom{N}{n}$

۱۴ . اگر در نمونه‌ای به‌حجم ۹ که به‌روش تصادفی سادهٔ بدون جایگذاری از جامعهٔ به‌حجم ۹۰ گرفته‌ایم، مجموع توان دوم مقادیر واحدهای نمونه برابر با ۴۴ و میانگین نمونه ۲ باشد، آن‌گاه برآورد واریانس میانگین نمونه برابر است با

الف) $\frac{۱}{۱۰}$ ب) ۱ ج) $\frac{۸}{۹۰}$ د) $\frac{۱}{۸}$

۱۵ . برای نمونهٔ تصادفی سادهٔ بدون جایگذاری از جامعه‌ای به‌حجم ۵۰، اگر s^2ی نمونه برابر با ۱۰۰ باشد آن‌گاه برآورد نااریب σ^2ی جامعه برابر است با

الف) ۵۰ ب) ۶۸ ج) ۱۰۰ د) ۲۵

۱۶ . نمونهٔ به‌حجم n از جامعهٔ به‌حجم N را نمونهٔ تصادفی سادهٔ بدون جایگذاری می‌نامند اگر

الف) احتمال عضویت همهٔ واحدهای جامعه در نمونه یکی باشد.

ب) احتمال وقوع همهٔ 2^N نمونهٔ ممکن یکی باشد.

ج) احتمال وقوع همهٔ $\binom{N}{n}$ نمونهٔ ممکن یکی باشد.

د) احتمال عضویت هر واحد جامعه در نمونه برابر با $\frac{1}{n}$ باشد.

۱۷ . میانگین یک جامعه برابر با ۲ است. حجم جامعه ۴۰ است. مجموع توان دوم مقادیر واحدهای جامعه ۲۰۰ است. با این داده‌ها، ضریب تغییرات جامعه برابر است با

الف) $\frac{۱}{۲}$ ب) $\frac{۱}{۴}$ ج) $\frac{۳}{۴}$ د) $\frac{۱}{۳}$

۱۸ . در یک نمونه‌گیری تصادفی ساده، برای تعیین برآورد نسبت یک مشخصه، نمونه‌ای به‌حجم ۲۰ به تصادف بدون جایگذاری گرفته‌ایم. اگر حجم جامعه ۲۰۰ و نسبت مشخصه در نمونه ۴ر۰ باشد، برآورد واریانس برآوردکنندهٔ نسبت، برابر است با

الف) $\frac{۱۰۸}{۹۵۰۰}$ ب) ۰۰۱۲ر۰ ج) $\frac{۹۵۰۰}{۱۰۸}$ د) ۵ر۰

۱۹ . از جامعه‌ای به‌حجم ۱۰ نمونه‌ای تصادفی بدون جایگذاری و به‌حجم ۵ گرفته‌ایم. اگر $\overline{Y}_n$ میانگین نمونه باشد، تعداد $\overline{Y}_n$های ممکن که شاید بعضی از آنها تکراری هم باشند برابر است با

الف) ۱۰۰۰۰۰ ب) $۵^{۱۰}$ ج) ۲۵۲ د) ۲

۲۰ . از روی یک نمونه‌گیری مقدماتی تصادفی سادهٔ بدون جایگذاری از جامعه‌ای به‌حجم ۱۰۰، برآورد S^2ی جامعه برابر با ۱۰۰ به‌دست آمده است. اگر بخواهیم برآورد واریانس میانگین یک نمونه تصادفی ساده بدون جایگذاری برابر با ۹ باشد باید حجم نمونه را برابر با

الف) ۱۰ ب) ۱۰۰ ج) ۲۰ د) ۱۲

انتخاب کنیم.