اساليب المعاينة

Sampling Method

محاضرات الدبلوم العالي في الاحصاء الحيوي/الكورس الاول

قسم الاحصاء/كلية الادارة والاقتصاد/الجامعة المستنصرية

اعداد

د.سهاد علي شهيد التميمي

2017/2016

المفردات

- الفصل الاول : مفاهيم وتعاريف إحصائية عامة مع توضيح طرائق المعاينة

Concepts and definitions of statistical sampling

1-1 General statistical terms

1-2 Types of samples

1-3 Error kinds of statistical sampling

1-4 Choosing a method of data collection

- الفصل الثاني: العينة العشوائية البسيطة

Simple Random Sampling

2-1 Introduction

2-1-1 Random Sampling

2-1-2 Nine drug addicts

2-2 Possible samples with or without replacement

2-3 Determination of sample size.

2-4 Find the error of random sampling.

Chapter three

Elementary probability and probability distribution

3-1 Introduction

3-2 Mutually exclusive events and additive law

3-4 Random variable and probability distribution

Chapter four

Demographic Method and health services statistics

4-1 Introduction

4-2 Source of demographic data

4-3 Vital statistics

4-3-1 Measures of Fertility and Mortality

4-4 Health services statistics

Chapter Five

Custer sample, systematic sample, stratified sample

5-1 Introduction

5-2 purpose of cluster sample.

5-3 systematic sample

5-4 stratified sample

5-5 examples of types of samples

Chapter six

Software Application on sampling method

6-1 Introduction

6-2 Spss

6-3 Stata

6-4 exercises

الفصل الاول

مفاهيم وتعاريف إحصائية عامـة

General Statistical Terms

1- الإحصاء Statistics

مجموعة الطرق والنظريات العلمية التي تعنى بجمع وعرض وتحليل البيانات الرقمية واستخدام نتائجها في أغراض التنبؤ أو التقدير أو التحقق أو اتخاذ القرار .

2 _ مصادر البيانات : Sources of Data

تنقسم المصادر التي تجمع منها البيانات الإحصائية ال لازمة للأبحاث والدراسات واتخاذ القرارات الى مصدرين هما :-

2 -1 المصادر التاريخية: Historical Sources

المصادر الرسمية للبيانات الإحصائية المحفوظة في السجلات والدوريات سواء سبق نشرها أو يتم نشرها بصفة دورية من قبل الأجهزة الإحصائية والهيئات المتخصصة في الدولة . ومن أمثلة المصادر التاريخية سجلات المواليد والوفيات والسجلات التجارية .

2 - 2 المصادر الميدانية Field Sources

ويقصد بها المصادر الأصلية التي تجمع منها البيانات الإحصائية ميدانيا سواء كان ذلك عن طريق المشاهدة المباشرة أو عن طريق استخدام استمارة احص ائية تعد لهذا الغرض ومن أمثلة المصادر الميدانية الأسر والمنشآت .

3 - عرض البيانات Data Representation

العملية التي تلي عملية جمع البيانات ومراجعتها حيث يتم عرض تلك البيانات بأشكال متعددة ، كعرضها في جداول خاصة أعدت بصورة مسبقة ، أو بأشكال هندسية أو رسوم بيانية ، وذلك لإجراء المقارنات السريعة بين مختلف اوجه الظاهرة المدروسة .

4 - وصف البيانات Data Description

بعد جمع البيانات وعرضها يهدف الإحصاء إلى دراسة الخصائص الأساسية للظاهرة المدروسة لوصفها وقياسها بمقاييس محدده تعبر عن هذه الخصائص ، ومن أهم هذه المقاييس المستخدمة لوصف مجموعة من البيانات مقاييس النزعة المركزية ومقاييس التشتت والالتواء والاعتدال .

5 - الوحدة الإحصائية Statistical Unit

الكيان أو الجزء الذي تجمع منه البيانات ، وتسمى وحدة العد التي تستخدم كأساس لجمع البيانات .

6 - المجتمع الإحصائي Statistical Population

جميع الوحدات الإحصائية التي يراد إجراء البحث الإحصائي عليها ، ومن الضروري تعريف هذه الوحدات بشكل واضح بحيث تجمعها صفة واحدة أو صفات مشتركة . ومعظم المجتمعات الإحصائية مؤلفة من وحدات إحصائية تتغير حسب الزمن (مجتمعات متجددة) ، وبعضها الأخر مجتمعات ثابتة لاتتغير حسب الزمن .

7 - الإطار Frame

قائمة أو سجل يشمل جميع وحدات المجتمع الإحصائي ، ويتضمن عادة أسماء وعناوين الوحدات الإحصائية وبعض المعلومات المتعلقة بها ، والإطار هو الدليل أو مجموعة الوثائق التي تساعدنا في الوصول إلى الوحدات الإحصائية لجمع البيانات عنها .

8 - أسلوب الحصر الشامل Census Methodology

أسلوب البحث الإحصائي الذي يدرس فيه حالة جميع وحدات المجتمع موضوع البحث دون استثناء ، وهذا يقتضي الوصول إلى كافة الوحدات الإحصائية لجمع البيانات عنها .

9 - العينة Sample

جزء من المجتمع الإحصائي يتم اختياره وفق أساليب المعاينة الإحصائية ويشترط أن تكون ممثلة للمجتمع الذي نقوم بدراسته ، ولكي تكون العينة ممثلة للمجتمع يجب أن تتضمن خصائص المجتمع بشكل يمكننا تعميم نتائجها لتقدير أهم معالم المجتمع الإحصائي .

10 - أنواع العينات Types of Samples

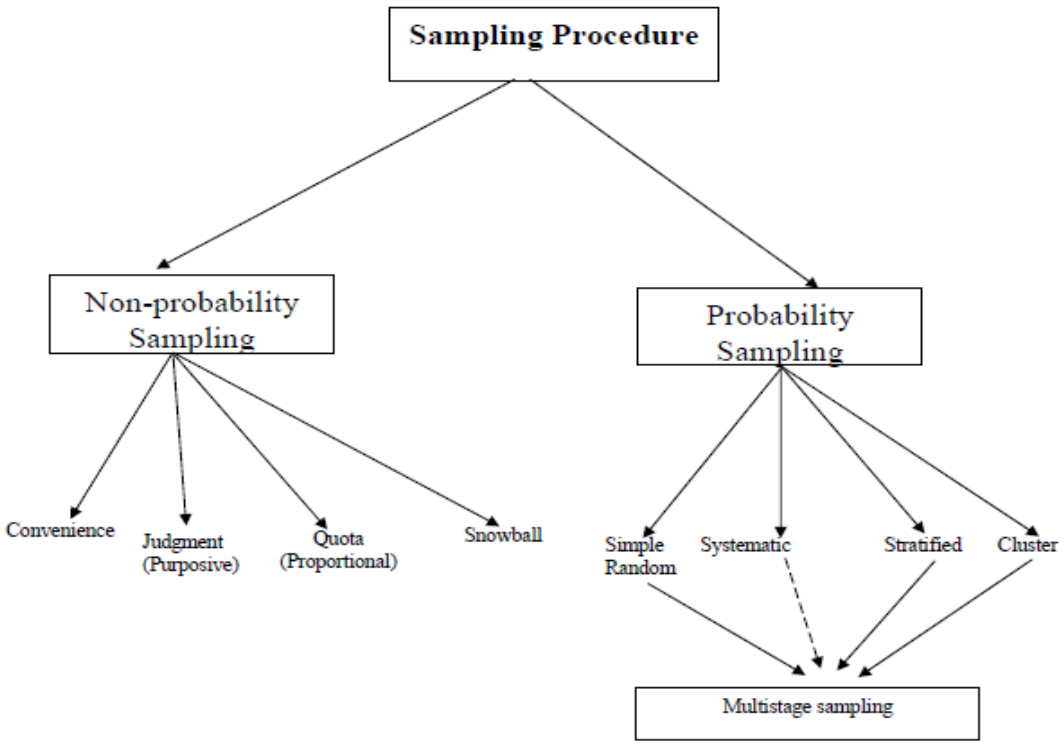
يمكن تقسيم العينات الى قسمين رئيسيين :

أولاً _ عينات احتمالية First – Probability Samples :

يتم سحبها على أساس قانون الاحتمالات حيث يتم سحب مفرداتها بشكل متتالي وباحتمال معروف ، ومن أنواعها العينة العشوائية البسيطة RandanSample والعينة الطبقية SlratifiedSample والعينة المنتظمة SystematicSample والعينة العنقودية ClusterSample .

ثانياً _ عينات غير احتمالية Second – Non–Probability Samples :

يتم اختيار وحداتها وفقا لمعايير يضعها الباحث دون التقيد بقوانين الاحتمالات ومنها العينة العمدية والعينة الحصصية quotaSample .



11 _ العينات الاحتمالية Probability Samples :

ومن أنواعها مايلي :

– العينة العشوائية البسيطة Simple Random Sample :

يتم اختيار مفرداتها على أساس إعطاء فرص متكافئة لجميع وحدات المجتمع ويعتمد كلية على عوامل الصدفة دون أية محاولة من الباحث للتحكم في الاختيار . وتستخدم عادة في حالة المجتمعات المتجانسة نسبيا ، والتي تقل فيها درجة التباين والاختلاف بين المفردات .

– العينة الطبقية Stratified Sample :

يقسم المجتمع موضوع الدراسة الى طبقات طبقا لمعايير معينة وتختلف هذه الطبقات فيما بينها من ناحية الخاصية التي نقيسها وفي كل طبقة تتشابه المفردات فيما بينها اكثر من تشابه العناصر داخل المجتمع بأكمله ، ويتم تطبيق أسلوب العينة العشوائية البسيطة في كل طبقة من الطبقات ، وعاده ماتستخدم العينة الطبقية في حالة المجتمعات غير المتجانسة .

– العينة المنتظمة Systematic Sample :

يتم اختيار مفردات العينة على أساس تقسيم المجتمع الى فترات متساوية عددها يساوي وحدات العينة المطلوبة ، ثم تحدد المفردة على أساس طريقة السحب العشوائي ، وبذلك يتعين موقع المفردة الأولى ثم نضيف الى هذا الرقم طول الفترة فيتحدد موقع المفردة الثانية ، ونضيف طول ا لفترة فنحصل على المفردة الثالثة ، وهكذا الى أن يتم سحب جميع مفردات العينة ، والعينات

المنتظمة شائعة الاستخدام ، نظرا لسهولتها ، ولكن الاختيار المنتظم قد يعرض العينة لبعض أخطاء التحيز ، إذا كان هناك توافق بين فترة الاختيار وأي اثر من آثار التغيرات الدورية في ترتيب الإطار المستخدم في سحب العينة .

- العينة العنقودية (المتعددة المراحل)Cluster (Multi-Stage) Sample :

لاختيار مفردات العينة المتعددة المراحل ، يقسم المجتمع إلى وحدات أولية ويتم اختيار عينة من هذه الوحدات كمرحلة أولى ثم تقسم كل وحدة من الوحدات الأولية المختارة إلى وحدات ثانوية تؤخذ منها عينة كمرحلة ثانية ، ثم تقسم كل وحدة من الوحدات الثانوية إلى وحدات اصغر تؤخذ منها كمرحلة ثالثة وهكذا حتى نحصل على حجم العينة اللازمة ، ويستخدم هذا النوع من العينات عندما يكون المجتمع كبيرا يصعب فيه إعداد إطار لجميع المفر دات ويقتصر تحضير الإطار للوحدات الثانوية فقط ، وهكذا نوفر الكثير من الجهود والتكاليف .

12 - العينات غير الاحتمالية:

وهي العينات التي لاتستخدم الطريقة العشوائية في الاختياربل تتاثر بالباحث وحكمه الشخصي ، هنالك دراسات يصعب تحديد المجتمع الأصلي لها مثل دراسة أحوال المدمنين أو المنحرفين أو المتهربين من الضرائب إن مثل هذه المجتمعات محددة وأفرادها ليسوا معروفين فلا نستطيع اخذ عينة عشوائية منهم بحيث تمثلهم بدقة ،فيعتمد الباحث إلى أسلوب العينة غير العشوائية ويختار عينة حسب معايير معينة يضعها الباحث ،فالباحث هنا يتدخل في اختيار العينة ويقرر من يختار ومن يهمل من المجتمع الأصلي للدراسة ،ولهذا الأسلوب ثلاثة أشكال من العينات:

-عينة الصدفة Accidental Sample او (Convenience sample)

يختار الباحث عددا من الافراد الذين قابلهم بالصدفة ،فإذا أراد الباحث إن يدرس موقف الرأي العام من قضية ما فانه يختار عددا من الناس يقابلهم بالصدفة في خلال ركوبه للسيارة أو وقوفه عند البائع أو في زاوية الطريق . ويؤخذ على هذه العينة انها لا يمكن إن تمثل المجتمع الأصلي بدقة ومن هنا يصعب تعميم نتائج البحث الذي يتناوله على المجتمع الأصلي كله.

عينة كرة الثلج Snowball sample:

حيث يعمد الباحث إلى انتقاء مبحوثين أوليين تبعا للمواصفات المطلوبة في دراسته، وبعد ذلك يطلب من المبحوث إرشاده إلى الشخص أو الوحدة الموالية التي تحمل نفس الخصائص المطلوبة ، وهكذا إلى أن يتم استيفاء التحقيق الميداني؛ وكأن العينة في هذه الحالة تبنى تدريجيا.

-العينة الحصصية Quota Sample

وهي عينة سهلة يمكن اختيارها بسرعة وسهولة حيث يقوم الباحث بتقسيم مجتمع الدراسة إلى فئات ،ثم يختار عدداً من أفراد كل فئة بحيث يتناسب مع حجم هذه الفئة ،فإذا أراد باحث إن يدرس موقف الرأي العام من قضية المحامين ،الأطباء... الخ،ثم يختار من كل فئة عددا من الأفراد ،إن هذه العينة تشبه العينة الطبقية العشوائية لكنها تختلف عنها في إن الباحث في العينة العشوائية لا يختار الأفراد كما يريد بينما في عينة الحصة يقوم الباحث بهذا الاختيار بنفسه

ودون إن يلزم نفسه بأية شروط فيتصل مع من يريد من الطلاب أ و المحامين أو العمال وبذلك لاتكون العينة ممثلة مجتمعها تمثيلا دقيقا.

-العينة الغرضية أو القصدية Purposive Sample

يقوم الباحث باختيار هذه العينة اختياراً حراً على أساس أنها تحقق أغراض الدراسة التي يقوم بها ،فإذا أراد باحث إن يدرس تاريخ التربية في العر اق فانه يختار عددا من المربين كبار السن كعينة قصدية تحقق أغراض دراسته ،انه يريد معلومات عن التربة القديمة في العراق وهؤلاء الأشخاص يحققون له هذا الغرض فلماذا لا يأخذهم كعينة؟ إذ ليس من الضروري إن تكون العينة ممثلة لأحد. فالباحث في هذه الحالة يقدر حاجته إلى المعلومات ويختار عينته بما يحقق له غرضه

12- الاستبانة الإحصائية Statistical Questionnaire :

الأداة الرئيسة التي يستخدمها الباحث لجمع البيانات من المبحوثين Respondents ، وتحتوي على الأسئلة التي تؤدي الإجابة عليها إلى الحصول على البيانات المطلوبة . ويجب الاهتمام والعناية بشدة بتصميم الاستبانة لكي تحقق الحصول على البيانات المطلوبة وبشكل سليم . وتقسم الاستبانة الإحصائية إلى نوعين هما :

_ استبانات فردية تخص وحدة إحصائية واحدة .

_ استبانات جماعية بحيث تخصص الاستبانة ال واحدة لعدد من الوحدات الإحصائية.

13- العينة الاستطلاعية (البحث التجريبي)Pilot Survey :

اختيار عدد من الوحدات الإحصائية وجمع البيانات عنها ، وتدوينها في استمارات متخصصة لهدف اختيار دقة تصميم الاستبانة ، والوقوف على الصعوبات التي قد يواجهها الباحثون عند تنفيذ البحث .

14- أنواع الخطاء الذي يتعرض له البحث الإحصائي

Errors Kinds of Statistical

هناك نوعين من الأخطاء التي قد يتعرض لها البحث الإحصائي وهما :

14 _ 1 الخطأ العشوائيRandom Error :

ويمكن التعرف عليه من مشاهدة انتشار نتائج البحث إذا تكرر إجراءه بنفس الأسلوب وتحت نفس الظروف . وهذا الخطاء لا يختفي عند استخدام أسلوب الحصر الشامل وذلك لانه ينتج عن اختلاف العدادين أو اختلاف الدافع الشخصي للإجابة على أسئلة البحث وفي معظم الأحيان يكون هذا الخطاء ضئيلا ويمكن قياسه ومعرفة حدوده .

ويتوقف مقدار هذا الخطأ على عام لين أساسيين هما مدى الاختلاف أو التباين بين وحدات المجتمع وحجم العينة بالنسبة للمجتمع الذي سحبت منه فكلما ازداد التباين بين وحدات المجتمع ازداد احتمال الوقوع في الخطأ العشوائي ، أما

بالنسبة لحجم العينة فكلما كبر حجم العينة انخفض احتمال الوقوع في هذا الخطأ .

14 _ 2 خطأ التحيز Bias Error :

وهو نوعان :

(أ) خطأ التحيز في التقدير :

وهو انحراف متوسط جميع التقديرات الممكنة لدليل المجتمع عن قيمته الحقيقية، ومن الصعب اكتشاف هذا الخطأ والتخلص منه إلا بأجراء تعديلات جذرية على تصميم البحث أو طريقة جمع البيانات أو تعديل النتائج .

(ب) خطأ التحيز في المعاينة :

وهو التحيز الذي يكون مقصودا ، وينشأ بسبب الإدلاء بمعلومات لاتطابق الواقع من قبل معطي البيانات أو نتيجة لتزوير بيانات الاستبانة من قبل الباحث ، أو مصممي البحث وفقا لميول أو أغراض مقصودة ، وأما أن يكون التحيز غير مقصود ، وهو الذي يتسرب إلى البحث لعدم فهم المبحوث للبيانات المطلوب تقديمها أو لعدم إتاحة الفرصة لتحضير إجابات صحيحة .

15- اختيار الوسيلة في جمع البيانات:

Choosing a method of data collection

- إذا كان الباحث بصدد اختيار العينة ، فإن عليه أن يعى تماماً أن هناك شرطاً رئيسياً يحكم قدرته على تعميم نتائجه على المجتمع الأصلى ، إنه التمثيل ، ويتطلب هذا توفر الشروط التالية:

- توافر كل صفات وخصائص المجتمع الأصلى فى العينة ، بحيث تكون نموذجاً مصغراً لهذا المجتمع ، وآنذاك نستطيع أن نقول : إن ما يصدق على هذا النموذج يصدق عِلى المجتمع الأصلى الذى اشتق منه.
- التناسب بين عدد أفراد العينة ، وعدد الأفراد الذين يشكلون المجتمع الأصلى ، فلا يكون المجتمع الأصلى طلاب المرحلة الثانوية مثلاً ، ويتخذ الباحث عينة عبارة عن فصل دراسى من إحدى المدارس الثانوية مكون من عشرين طالباً.
- منح جميع أفراد المجتمع الأصلى فرصة متكافئة لأن يتم اختيارهم للانضمام للعينة ، بمعنى آخر موضعية الاختيار وعدم التحيز لفرد معين أ و فئة معينة دون غيرها.
- نوع المجتمع الأصلى : فإذا كان هذا المجتمع متجانساً فإن الباحث يكتفى بدراسة عينة صغيرة منه ، ويعمم النتائج على هذا المجتمع ، أما إذا كان هذا المجتمع متبايناً غير متجانس ويحتوى مجموعات فرعية كثيرة فلابد للعينة أن تكون كبيرة لاستيعاب هذا التباين .
- كلما كان الإنحراف المعيارى صغيراً كلما قل تشتت "تباين" الدرجات وزاد تجانسها . وإذا زاد الإنحراف المعيارى زاد تشتت الدرجات وقل تجانسها.

- إذا كان الباحث يتوقع الحصول على فروق ضيئلة ، أو علاقات غير قوية ، يجب أن يجعل العينة كبيرة لتتضح هذه الفروق ، مثال ذلك يتوقع من التدريب ان يحدث تغيرات بسيطة فى تحصيل الطلاب ، لكن إذا كانت هذه التغيرات ذات قيمة للباحث ، فإنه يتحتم عليه تجنب العينات الصغيرة حتى لا تطمس هذه التغيرات.
- حجم العينة الصغير مقبول فى الدراسات الاستطلاعية ، وذلك لأن البلحث يتحمل هامش كبير نسبياً من الخطأ فى النتائج . إلا أنه فى الدراسات التى يترتب عليه توزيع الأفراد على مجموعات أو اتخاذ قرار فمن الأفضل وجود عينة كبيرة بشكل كاف لتقليل الخطأ.
- إذا لم تكن أدوات جمع البيانات دقيقة أو ثابتة بدرجة مرتفعة يفضل استخدام عينة كبيرة لتعويض خطأ جمع البيانات .
- يتأثر حجم العينة بنوع الأداة المستخدمة فى جمع البيانات (المقابلة ، والملاحظة ، والاختبارات الفردية تستلزم عينات صغيرة . أما الاختبارات الجمعية والاستبيانات يمكن استخدام عينات كبيرة).
- تزداد دقة النتائج ويصبح من الممكن التعميم منها على المجتمع كلما زاد حجم العينة . ولكن يلاحظ أن هناك حداً امثل لحجم العينة إذا تخطاه الباحث فإنه لن يستفيد كثيراً من زيادة عدد الأفراد فى عينته.
- عند استخدام الانحدار المتعدد أو الاختبارات المماثلة له فان حجم العينة يجب أن يكون عشر أضعاف متغيرارت الدارسة .مثلا إذا احتوت الدارسة على 6 متغيرات لإجراء التحليل عليها فانه يفضل ألا يقل حجم العينة عن 60 مفردة

والجدول التالي يبين حجم العينة المناسب عند مستويات مختلفة من مجتمع الدراسة الأصلي:

حجم المجتمع الأصلي	حجم العينة المناسب	حجم المجتمع الأصلي	حجم العينة المناسب
10	10	550	226
30	28	650	242
70	59	900	269
110	86	1100	285
170	118	2000	322
210	136	6000	361
250	152	15000	375
360	186	75000	382
420	201	1000000	384

Source: Uma Sekaran, 1992.

الفصل الثاني
العينة العشوائية البسيطة
Simple Random Sampling

2-1 Introduction

Everyone mentions simple random sampling, but few use this method for population based surveys. Rapid surveys are no exception, since they too use a more complex sampling scheme. So why should we be concerned with simple random sampling? The main reason is to learn the theory of sampling. Simple random sampling is the basic selection process of sampling and is easiest to understand.

When the true value in a population is estimated with a sample of persons, things get more complicated. Rather then just the mean or proportion, we need to derive the standard error for the variable of interest, used to construct a confidence interval. This chapter will focus on simple random sampling or persons or households, done both with and without replacement, and present how to derive the standard error for equal interval variables, binomial variables, and ratios of two variables. The latter, as described earlier, is commonly used in rapid surveys and is termed a *ratio estimator*. What appears to be a proportion, may actually be a ratio estimator, with its own formula for the mean and standard error.

2.1.1 Random sampling

Subjects in the population are sampled by a random process, using either a random number generator or a random number table, so that each person remaining in the population has the same probability of being selected for the sample. The process for selecting a random sample is shown in Figure 2-1.

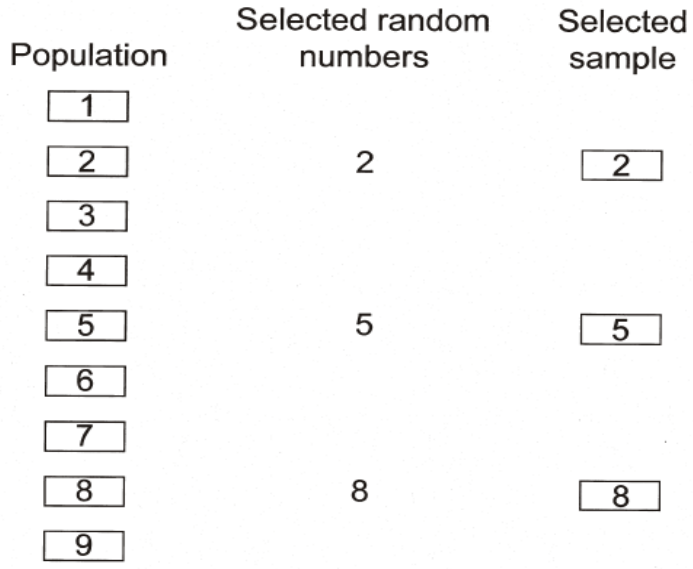
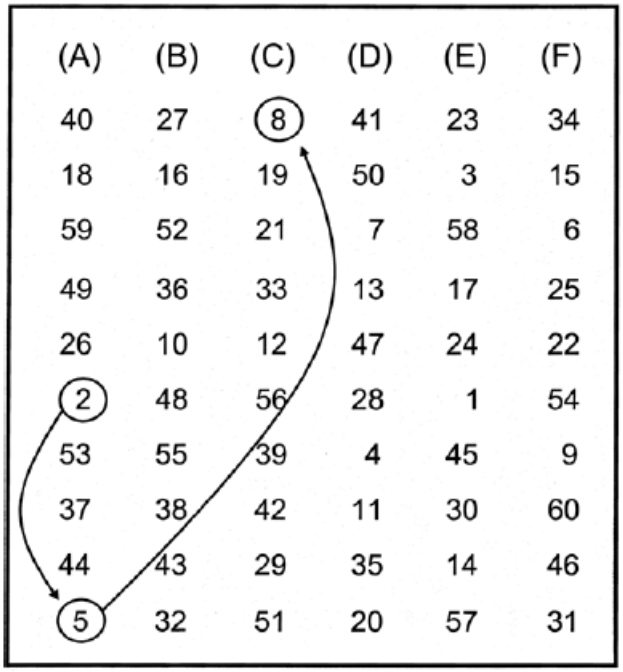


Figure 2-1. Random sample of three units from a population of nine units.

The population to be sampled is comprised of nine units, listed in consecutive order from one to nine. The intent is to randomly sample three of the nine units. To do so, three random numbers need to be selected from a random number table, as found in most statistics texts and presented in Figure 2-2. The random number table consists of six columns of two-digit non-repeatable numbers listed in random order.



(A)	(B)	(C)	(D)	(E)	(F)
40	27	8	41	23	34
18	16	19	50	3	15
59	52	21	7	58	6
49	36	33	13	17	25
26	10	12	47	24	22
2	48	56	28	1	54
53	55	39	4	45	9
37	38	42	11	30	60
44	43	29	35	14	46
5	32	51	20	57	31

Figure 2-2

The intent is to sample three numbers between 1 and 9, the total number in the population. Starting at the top of column A and reading down, two numbers are selected, 2 and 5. In column B there are no numbers between 1 and 9. In column C the first random number in the appropriate interval is 8.

Thus in our example, the randomly selected numbers are 2, 5 and 8 used to randomly sample the subjects in Figure 2-1. Since the random numbers are mutually exclusive (i.e., there are no duplicates), each person with the illustrated method is only sampled once. As described later in this chapter, such selection is sampling without replacement.

Random sampling assumes that the units to be sampled are included in a list, also termed a sampling *frame*. This list should be numbered in sequential order from one to the total number of units in the population. Because it may be time-consuming and very expensive to make a list of the population, rapid surveys feature a more complex sampling strategy that does not require a complete listing. Here, however, every member of the population to be sampled is listed

2.1.2 Nine drug addicts

A population of nine drug addicts is featured to explain the concepts of simple random sampling. All nine addicts have injected heroin into their veins many times during the past weeks, and have often shared needles and injection equipment with colleagues. Three of the nine addicts are now infected with the human immunodeficiency virus (HIV). To be derived are the proportion who are HIV infected (a binomial variable), the mean number of intravenous injections (IV) and shared IV injections during the past two weeks (both equal interval variables), and the proportion of total IV injections that were shared with other addicts. This latter proportion is a ratio of two variables and, as you will learn, is termed a *ratio estimator*.

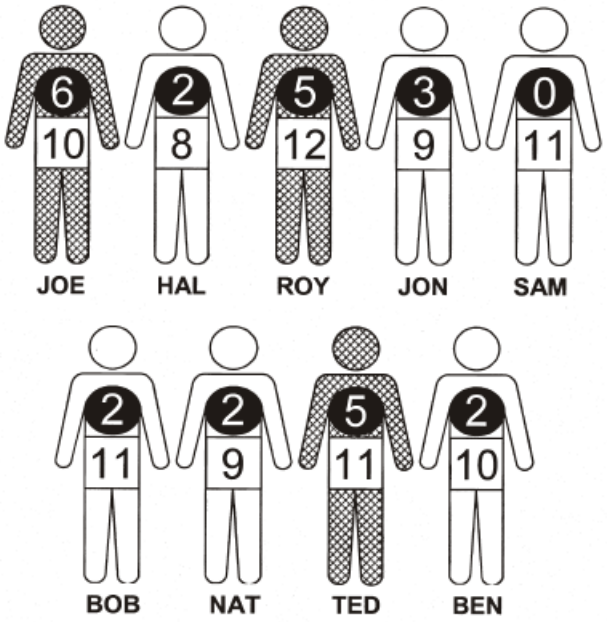


Figure 2-3. Population of nine intravenous drug addicts

The total population of nine drug addicts is seen in Figure 2-3. Names of the nine male addicts are listed below each figure. The three who are infected with HIV are shown as cross-hatched figures. Each has intravenously injected a narcotic drug eight or more times during the past

two weeks. The number of injections is shown in the white box at the midpoint of each addict. With one exception, some of the intravenous injections were shared with other addicts; the exact number is shown in Figure 2-3 as a white number in a black circle.

Our intention is to sample three addicts from the population of nine, assuming that the entire population cannot be studied. To provide an unbiased view of the population, the sample mean should on average equal the population mean, and the sample variance should on average equal the population variance, corrected for the number of people in the sample. When this occurs, we can use various statistical measures to comment about the truthfulness of the sample findings. To illustrate this process, we start with the end objective, namely the assessment of the population mean and variance.

Population Mean. For total intravenous drug injections, the mean in the population is derived using Formula 1

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \quad \ldots(1)$$

Where *Xi* is the total injections for each of the *i* addicts in the population and *N* is the total number of addicts. Thus, the mean number of intravenous drug injections in the population shown in Figure 2-3 is:

$$\bar{X} = \frac{10+8+12+9+11+11+9+11+10}{9} = \frac{91}{9} = 10.1$$

Or 10.1 intravenous drug injections per addict.

Population Variance. Formula 2 is used to calculate the variance for the number of intravenous drug injections in the population of nine drug addicts.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} \quad \ldots(2)$$

Where σ^2 is the Greek symbol for the population variance, *Xi* and *N* are as defined in Formula 1 and $\bar{X}$ is the mean number of intravenous drug injections per addict in the population . Using Formula 2, the variance in the population is

$$\sigma^2 = \frac{(10-10.1)^2 + (8-10.1)^2 \ldots (11-10.1)^2 + (10-10.1)^2}{9} = 1.43$$

Sample Mean. Since the intent is to make a statement about the total population of nine addicts, a sample of three addicts will be drawn, and their measurements will be used to represent the group.

The three will be selected by simple random sampling. The mean for a sample is derived using Formula 3.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad ...(3)$$

where x_i is the number of intravenous injections in each sampled person and n is the number of sampled persons. For example, assume that Roy-Jon-Ben is the sample. Roy had 12 intravenous drug injections during the past two weeks (see Figure 2-3), Jon had 9 injections and Ben had 10 injections. Using Formula 3

$$\bar{x} = \frac{12+9+10}{3} = 10.3$$

the sample estimate of the mean number of injections in the population (seen previously as 10.1) is 10.3.

Sample Variance. The variance of the sample is used to estimate the variance in the population and for statistical tests. Formula 4 is the standard variance formula for a sample.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad ...(4)$$

where s^2 is the symbol for the sample variance, x_i is the number of intravenous injections for each of the i addicts in the sample and is the mean intravenous drug injections during the prior week in the sample. For the sample Roy-Jon-Ben with a mean of 10.3, the variance is

$$s^2 = \frac{(12-10.3)^2 + (9-10.3)^2 + (10-10.3)^2}{3-1} = 2.33$$

2-1-3 Samples with or without replacement:

There are two ways to draw a sample, with or without replacement. With replacement means that once a person is selection to be in a sample, that person is placed back in the population to possibly be sampled again. Without replacement means that once an individual is sampled, that person is not placed back in the population for re-sampling.

An example of these procedures is shown in Figure 3-4 for the selection of three addicts from a population of nine

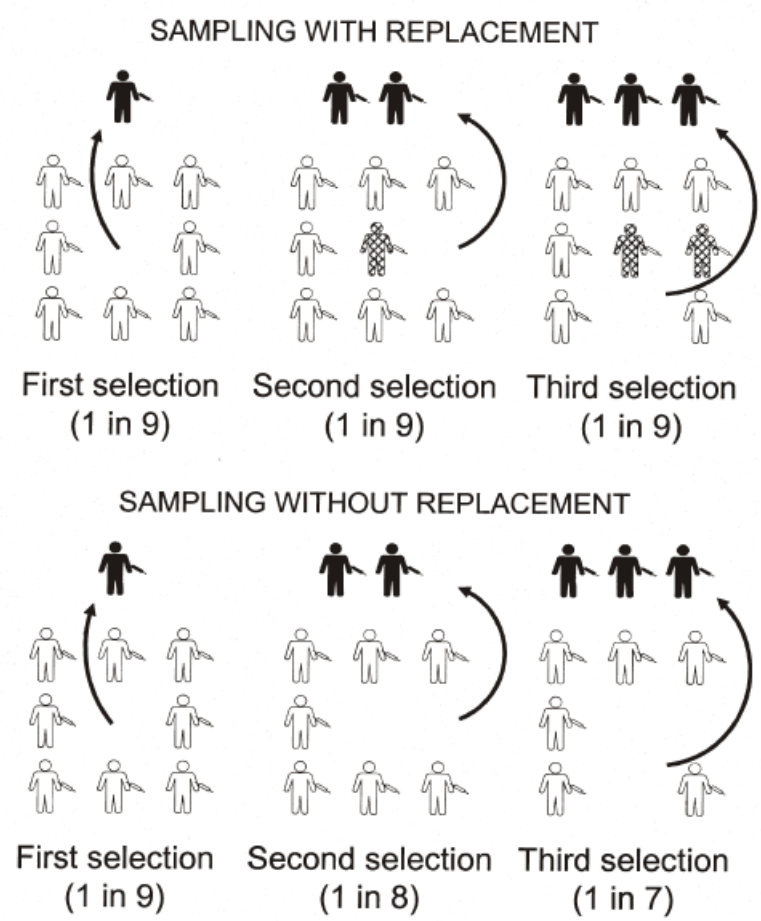


Figure 2-4. Sample of three addicts from a population of nine addicts.

In sampling *with replacement* (Figure 2-4, top), all nine addicts have the same probability of being selected (i.e., 1 in 9) at steps one, two and three, since the selected addict is placed back into the population before each step. With this form of sampling, the same person could be sampled multiple times. In the extreme, the sample of three addicts could be one person selected three times.

2-1-4 Determination of sample size

In sampling *without replacement* (WOR) the selection process is the same as at step one) that is each addict in the population has the same probability of being selected (Figure 2-4, bottom). At step two, however,

the situation changes . Once the first addict is chosen, he is not placed back in the population. Thus at step two, the second addict to be sampled comes from the remaining eight addicts in the population, all of whom have the same probability of being selected (i.e., 1 in 8). At the third step, the selection is derived from a population of seven addicts, with each addict having a probability of 1 in 7 of being selected. Once the steps are completed, the sample contains three different addicts.

When drawing a sample from a population, there are many different combinations of people that could be selected. Formula 3.6 is used to derive the number of possible samples drawn *with replacement*,

$$N^n$$

where N is the number in the total population and n is the number of units being sampled. For example when selecting three persons from the population of nine addicts shown in Figure 2-3, the sample could have been Joe-Jon-Hall, or Sam-Bob-Nat, or Roy-Sam-Ben, or any of many other combinations. To be exact, in sampling *with replacement* from the population shown in Figure 2.3, there are

$$N^n = 9^3 = 729$$

or 729 different combinations of three addicts that could have been selected.

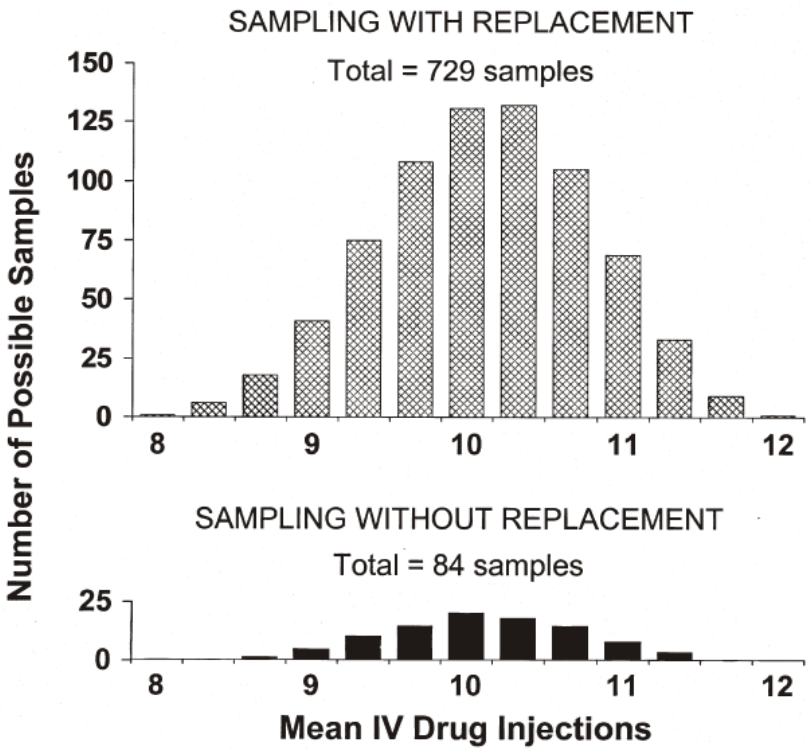


Figure 2-5. Distribution of all possible sample means *with* and *without replacement* (actual scale).

The frequency distribution of the mean number of IV drug injections of the 729 possible samples selected *with replacement* is shown in the top section of Figure 2-5. Notice that the distribution has a bell shape, similar to a normal curve. There are three notable features of these 729 possible samples.

The average variance of the 729 possible samples of three selected *with replacement* is equal to the population variance of the nine drug addicts (see Formula 2), as shown in Formula 7.

$$\sigma^2 = \frac{\sum_{i=1}^{729} s_i^2}{729} = 1.43 \qquad ...(7)$$

For the 729 possible samples, the average variance of the mean for a sample of three from an underlying population of nine is shown in Formula 8.

$$v(\bar{x}) = \frac{\sum_{i=1}^{729} (\bar{x}_i - \bar{X})^2}{729} = \frac{\sigma^2}{n} = \frac{1.43}{3} = 0.48 \qquad ...(8)$$

Without Replacement. In the realistic world of sampling, subjects are typically not included in the sample more than once. Also, the order in which subjects are selected for a survey is not important (that is, Roy-Sam-Ben is considered the same as Sam-Ben-Roy). All that matters is if the subject is in or out of the sample. Hence in most surveys, samples are selected *disregarding order* and *without replacement*. But does sampling *without replacement* provide unbiased estimators of the population mean and variance? The answer is "yes," but needing some additional modifications, to be presented next. Formula 9 is used to calculated the number of possible samples that can be drawn *without replacement, disregarding order,*

$$\frac{N!}{n!\ (N-n)!} \qquad ...(9)$$

In our example, we are selecting *without replacement* and *disregarding order* a sample of three addicts from a population of nine addicts (see Figure 2-3). Using Formula 9, we find there are

$$\frac{9!}{3!\,(9-3)!} = \frac{9 \times 8 \times 7 \times 6!}{(3 \times 2 \times 1) \times 6!} = \frac{504 \times 6!}{6 \times 6!} = \frac{504}{6} = 84$$

or 84 possible samples. Fortunately when using Formula 9, all factorial numbers do not have to be multiplied. For example, the *9!* in the numerator can be converted to *9 x 8 x 7 x 6!*, and the *3! * (9-3)!* in the denominator can be converted to 3 x *2 x 1 x 6!*. By dividing *6!* in the numerator by *6!* in the denominator to get 1, the formula is reduced to *9 x 8 x 7* divided by 3 x *2 x 1* or 84 possible samples.

The distribution of all possible sample means for the 84 samples selected *with replacement, disregarding order* in shown in the bottom section of Figure 2-5, below the distribution of the 729 possible sample means selected *with replacement*. Are the two distributions similar? It is hard to tell since the scale does not permit an easy visual comparison. Figure 2-6 shows the same two distributions, but as a percentage of the total number of possible samples (i.e., 729 *with replacement* and 84 *without replacement*).

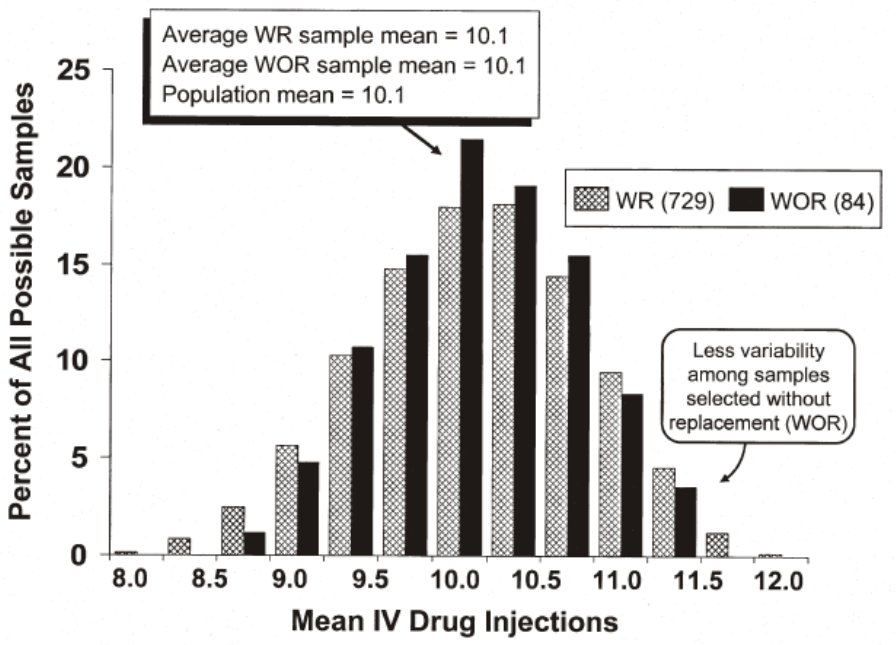


Figure 2-6. Distribution of all possible sample means *with* and *without replacement* (percentage scale).

There are two things to notice. First, the mean of all possible samples selected *with replacement* (i.e., 10.1) is equal to the mean of all samples selected *without replacement*, and both sample means are equal to the population mean. Thus, the sample mean on average remains

an unbiased estimator of the population mean when sampling *without replacement*. Second, the percentage distributions of those selected with and without replacement are similar in shape, but there are fewer outlying samples among those sampled *without replacement*. That is, there is less variability among the 84 possible samples selected *without replacement* than the 729 possible samples selected *with replacement*. The reduced variability in sampling *without replacement* is addressed in two ways, namely with a change in the variance formula for the population variance and in the addition of a finite population correction factor (FPC).

First, different from Formula 2, the population variance that is being estimated by the sample variance when sampling *without replacement* has a different denominator (N-1), as shown in Formula 10.

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1} \quad ...(10)$$

where S^2 is the modified population variance and Xi , N and are as defined previously. For the population of nine drug addicts, the modified variance is

$$S^2 = \frac{(10-10.1)^2 + (8-10.1)^2 \ldots (11-10.1)^2 + (10-10.1)^2}{8} = 1.61$$

When sampling *without replacement* the average variance of all 84 possible samples is equal to the modified population variance (see Formula 11).

$$\frac{\sum_{i=1}^{84} s_i^2}{84} = 1.61 = S^2 \quad ...(11)$$

sample i, with i going from 1 to 84, the total number of possible samples when selecting three from nine *without replacement*.

Second, the variance of the sample mean of all 84 possible samples when sampling *without replacement* is equal to the modified population variance divided by the sample size times a correction factor that accounts for the shrinkage in variance. This correction factor, termed the *finite population correction* (FPC) is shown in Formula 12.

$$FPC = \frac{N-n}{N} = 1 - \frac{n}{N} \quad \text{...(12)}$$

where N is the size of the population and n is the size of the sample. In samples where the sample size is large in relation to the population (an example being a sample of three from a population of nine), the FPC reflects the reduction in variance that occurs when sampling *without replacement* (i.e., with 84 possible samples in the example) compared to sampling *with replacement* (i.e., with 729 possible samples in the example). This reduction in variability when sampling *without replacement* was observed in Figure 2.6, and in the comment that there were fewer outliers in the *without replacement* group.

For the 84 possible samples, the average variance of the mean for a sample of three from an underlying population of nine is shown in Formula 13.

$$v(\bar{x}) = \frac{\sum_{i=1}^{84} (\bar{x}_i - \bar{X})^2}{84} = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{1.61}{3}\left(1 - \frac{3}{9}\right) = 0.36 \quad \text{...(13)}$$

Notice that n/N is the fraction of the population that is sampled. Therefore the FPC is often described by sampling specialists as "one minus the sampling fraction." Notice also that the variance of the average samples mean is 0.36 for sampling *without replacement* compared to 0.48 (see Formula 3.8) when sampling *with replacement*, resulting in smaller estimates of sampling error and greater efficiency in the sampling process when the sampling fraction is large. Finally, note that if the sampling fraction is very small, as occurs in typical rapid surveys of few persons drawn from a large population, then the finite population FPC term reduces to approximately 1, and is no longer needed

(Notes): We start with a short description of a number of important population parameters. Given a specific target variable we distinguish the population total Y, the population mean $\bar{Y}$, the population variance σ_y^2 , the adjusted population variance S_y^2 , and the population coefficient of variation CV_y. The adjusted population variance is often used to simplify the SRSWOR formulas.

Simple random sampling without replacement (SRSWOR) is the most familiar sampling design. This kind of sampling is referred to as simple because it involves drawing from the entire population.

Alternatively, simple random sampling can be carried out with replacement (SRSWR)

The rather formal definition of SRSWOR is as follows. Consider a population U of N elements, or U = {1,2,L,N}. SRSWOR is a method of selecting n elements out of U such that all possible subsets of U of size n have the same probability of being drawn as a sample.

Note that there are $\binom{N}{n}$ possible subsets of U of size n. In practice SRSWOR can involve successively selecting random numbers between 1 and N . The adjusted population variance is often used to simplify the SRSWOR formulas Moreover, Yk stands for the value of the target variable of element k, k=1.... N . The stated parameters are defined as follows:

$$Y = Y_1 + \cdots + Y_N = \sum_{k=1}^{N} Y_k$$

$$\bar{Y} = \frac{1}{N} Y = \frac{1}{N} \sum_{k=1}^{N} Y_k$$

$$\sigma_y^2 = \frac{1}{N} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$$

$$S_y^2 = \frac{N}{N-1} \sigma_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$$

$$CV_y = \frac{S_y}{\bar{Y}} \quad .$$

The n (i.e. the sample size) observations of the target variable in the sample are denoted in small letters y1 ..... yn , The sample mean $\bar{y}_s$, the sample variance S_y^2 , and the sample coefficient of variation cv_y are the three most important sample parameters. These three sample parameters are defined as follows:

$$\bar{y}_s = \frac{1}{n} \sum_{k=1}^{n} y_k$$

$$s_y^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y}_s)^2$$

$$cv_y = \frac{s_y}{\bar{y}_s} \quad .$$

Example There are 169 industrial establishments employing 20 or more software testers in IBM. The following are the employment figures based on a 1-in-5 systematic sample

35, 88, 35, 36 ,156, 25, 24, 237, 80 ,468 ,22,139, 163 ,37 ,37, 27, 25, 26, 38, 24, 62, 331, 28 ,31 ,81, ,121 ,49, 23, 34 ,23, 22, 53, 50 ,50

N = 169
n = 34
The average of samples,
$\acute{y} = 1/n \Sigma y_i = (35 + 88 + 35 + \ldots\ldots 53 + 50 + 50) /34 = 78.83$

$[1/ (N -1)] * \Sigma S (y_i - \acute{y})^2 = [1/168] * ((35 - 78.83)^2 + (88 - 78.83)^2 + \ldots\ldots (50 - 78.83)^2) = 309,795 /34 = 9387.73$

Variance of the estimate, $V (\acute{y}) = 1/n (1 - n/N) [1/ (N -1)] * \Sigma S (y_i - \acute{y})^2 = 1/34 * (1 - 34/169) * 9387.73 = 220.89$

Standard Error = $\sqrt{(V (\acute{y}))} = \sqrt{220.89} = 14.86$

Coefficient of Variation = Standard Error/ $\acute{y} * 100 \% = 18.9 \%$

مثال : إذا كان عدد البيوت البلاستيكية الز ا رعية (دفيئة) التي يزرع بيا محصول البندورة في محافظة كربلاء 450 دفيئة ، كيف يمكن أخذ عينة عشوائية حجميا 6% ، تمثل المجتمع الإحصائي بشكل سليم .

الحل:

عدد افراد العينة = (عدد افراد المجتمع)×(حجم العينة) / 100 = 6 × 450 / 100 = 27

العينة عشوائية بسيطة

الرقم	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
العينة	111	212	104	367	172	242	169	294	296	126	145	170	91	84
الرقم	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	
العينة	100	355	74	353	109	407	432	159	384	440	84	351	259	

2-4 Find the error sampling.

Error can occur during the sampling process. **Sampling error** can include both systematic sampling error and random sampling error. **Systematic sampling error** is the fault of the investigation, but **random sampling error** is not.

When errors are systematic, they bias the sample in one direction. Under these circumstances, the sample does not truly represent the population of interest. Systematic error occurs when the sample is not drawn properly, as in the poll conducted by *Literary Digest* magazine.

It can also occur if names are dropped from the sample list because some individuals were difficult to locate or uncooperative. Individuals dropped from the sample could be different from those retained.

Random sampling error, as contrasted to systematic sampling error, is often referred to as *chance error*. Purely by chance, samples drawn from the same population will rarely provide identical estimates of the population parameter of interest. These estimates will vary from sample to sample.

For example, if you were to flip 100 unbiased coins, you would not be surprised if you obtained 55 heads on one trial, 49 on another, 52 on a third, and so on. Thus, some samples will, by chance, provide better estimates of the parameter than others.

Standard Error ($s_{\bar{y}}$)	
Sampling with Replacement (or from infinite population)	Sampling without Replacement from a Finite Population
$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}}$	$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}\left(\frac{N-n}{N}\right)}$
where, $n = sample\ size$ and $N = population\ size$	