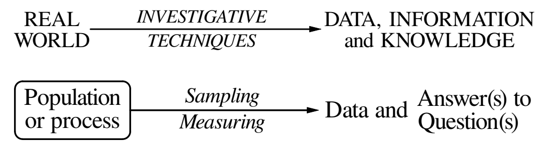
University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Populations and Processes
and Their Broader Implications

1. Background – What is Statistics About?

As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* (or empirical problem solving) of the real world, which means investigating some population or process on the basis of *data* to *answer* one or more *questions.* For this Highlight #94 with a specific focus, only the investigative techniques of *sampling* and *measuring* are shown in the lower schema.



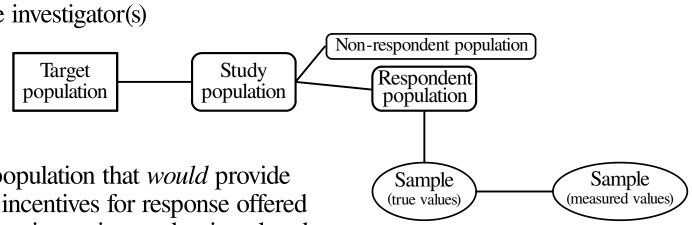
2. Four Populations and the Sample – Some Distinctions

* **Population:** a well-defined (finite) group of *elements* ***other than*** the sample. [An *infinite* 'population' is a (sometimes useful) ***model***.]
* **Element:** the entities that make up a population; for example, a person is an element of the population of Canadians, but we recognize that many populations in data-based investigating have non-human or *in*animate elements.
* **Unit:** the entities *selected* for the sample; a unit may be one element (*e.g.*, a person) or *more than* one (*e.g.*, a household).

The element-unit distinction is pursued briefly in Note 2 overleaf on page HL94.2.

* **Sampling:** includes the processes of *selecting* and *estimating.*
* **Measuring:** the process used to determine the value of a variate.

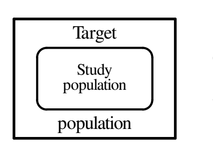
The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating.



* **Target population**: the group of elements to which the investigator(s) want Answer(s) to the Question(s) to apply.
* **Study population**: a group of elements *available* to an investigation.
* **Respondent population:** those elements of the study population that *would* provide the data requested under the incentives for response offered in the investigation [such incentives arise predominantly when the elements are people, but missing data may also arise when elements are *in*animate].
* **Non-respondent population:** those elements of the study population that would *not* provide the data requested under the incentives for response offered in the investigation – see also Note 1 overleaf on page HL94.2.
* **Sample**: the group of units selected from the respondent population *actually used* in an investigation – the sample is a *sub*set of the respondent population (as the vertical line in the schema above reminds us).
  - **Selection:** the units selected from the study population and which comprise the respondents (the sample) *and* the non-respondents – see Note 1 overleaf on page HL94.2. ['Selection' is terminology *specific* to these Materials.]
  - **Census:** using *all* the respondent population elements in an investigation; *i.e.*, in a census, the 'sample' size is n = N.
* **Variate:** a characteristic associated with each *element* of a population/process.
  - **Response variate (Y):** a variate defined in the Formulation stage of the FDEAC cycle; an Answer gives some attribute(s) of the response variate over the target population/process. [See Table HL94.2 overleaf on page HL94.2.]
  - **Explanatory variate (Z):** a variate, defined in the Formulation stage of the FDEAC cycle, that accounts, at least in part, for changes from element to element in the value of a response variate.
  - **Focal (explanatory) variate (X):** for a Question with a *causative* aspect, the focal variate is the *explanatory* variate whose relationship to the *response* variate is involved in the Answer(s) to the Question(s).
* **Attribute**: a quantity defined as a function of response (and, perhaps, explanatory) variates over any of the five groups of elements defined above.

  Familiar (simple) attributes are averages, proportions, medians, totals and s.d.s; the importance of attributes is that:
  + Answer(s) to Questions(s) are usually given in terms of attributes, often their values;
  + five of the six categories of *error* are defined in terms of attributes – see, for example, the upper half of page HL18.4 in Statistical Highlight #18.

Target elements and study elements are often the same but *may* differ; for instance, when assessing drug efficacy and side effects using laboratory animals, target elements are humans but study elements are laboratory animals (see the newspaper article *Miracle cure that kills claims fifth victim* in Statistical Highlight #22). The possibility of different target and study elements may be obcured by a (misleading) diagram like the one at the right which shows the study population as a *subset* of the target population.



- Such a diagram may also (mistakenly) show the *sample* as a subset of the study (rather than the respondent) population; correctly showing relevant relationships requires a more extensive schema, like the one above at the centre right, for instance.

(*continued overleaf*)

University of Waterloo W. H. Cherry

3. Processes

To answer some types of Question in statistics, instead of a population, we start with a *process*; we distinguish two cases:

∗ **Process:** ● a set of *operations* that produce or affect elements or units, OR:
● the *flow* of an an entity (like water or electrons).

The first case arises when the Question is about improving a manufacturing or service-delivery process; we quantify the performance of the process by measuring variate values on the elements or units it produces or affects.

○ The target process is typically the process now and into the future for as long as the current (or improved) implementation of the process operates.

The second case arises when the Question is about an entity that flows, like water in a river or electrons in a circuit or network; we quantify characteristic(s) of such processes by measuring variate values on the entity that flows.

○ The target process is typically the process over a defined period of time.

An investigation with a target *population* will have a study *population*; an investigation with a target *process* that is a set of operations will also have a study *population* – the available elements or units produced or affected by the process. An investigation with a target *process* that flows will have a study *process*, usually the target process over a restricted time period (see Table HL94.1 at the right).

Table HL94.1
Populations and Processes

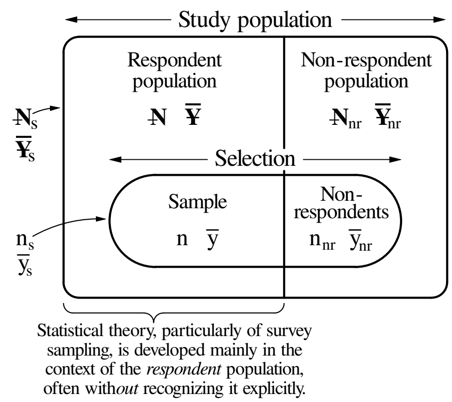
Target population	----------------►	Study population
Target process: operations	------►	Study population
Target process: flow	------------►	Study process

NOTES: 1. As indicated in the diagram at the right, *we* consider the *study* population to be made up of the *respondent* and *non-respondent* populations. The set of units selected from the study population is the **selection**, and comprises the *sample* (from the respondent population) and the *non-respondents* (from the *non*-respondent population). The diagram has *two* categories of symbols:

– the $\mathbf{N}$s and ns refer to *numbers of elements or units*;
– the $\bar{\mathbf{Y}}$s and $\bar{y}$s are *averages* of a response variate $\mathbf{Y}$ of the elements or units.

The relationships among the numbers of elements or units are:

Study population	=	Respondent population	+	Non-respondent population
$\mathbf{N}_s$	=	$\mathbf{N}$	+	$\mathbf{N}_{nr}$
Selection	=	Sample	+	Non-respondents
n_s	=	n	+	n_{nr}



2. In the STAT 231 Course Materials, the element-unit distinction is largely ignored and 'unit' is used throughout (for consistency with the STAT 231 Couse Notes). In the STAT 220 Materials, attention is generally restricted to units which are elements so the distinction does not arise but, more broadly, when units are *groups* of elements (as in *cluster* sampling, for instance), the distinction is essential; the matter is illustrated in more detail in Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77.
   - Materials like Statistical Highlight #77, concerned with survey *sampling*, refer in most places to units, but respondent population attributes of interest [like $\mathbf{N}$ (size), $\bar{\mathbf{Y}}$ (average), ${}_T\mathbf{Y}$ (total) and $\mathbf{S}$ (standard deviation)] refer to *elements.*

3. The term 'process', as discussed above in Section 3, is to be set against 'population' as *we* use these two terms in a statistical context. In ordinary English, 'process' also encompasses *a series of actions that bring about an end or result*; this meaning (which covers the first case of 'process' defined above) is why we *also* refer to selecting, measuring and estimating as processes – sampling is then a process too because it includes selecting and estimating.

Table HL94.2: The FDEAC cycle: a *structured process* for data-based investigating

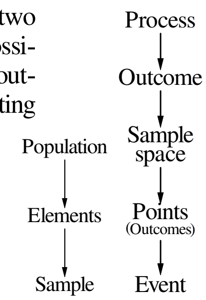
Stage	Formulation stage	Design stage	Execution stage	Analysis stage	Conclusion stage
Input	***Question(s)***	***clear Question(s)***	***a Plan***	***Data***	***Information***
Components	Target element Target population/process Variates: ● Response ● Explanatory Attributes Fishbone diagram Aspect: ● Descriptive ● Causative	Study element/unit Study population/process Respondent population/process Refine response variate(s) Deal with explanatory variates Protocol for: ● Selecting units ● Choosing groups ● Setting levels Measuring process(es) Plan for the: ● Execution stage ● Analysis stage	Execute the Plan Monitor the data Examine the data Store the data	Informal analysis: ● Numerical attributes ● Graphical attributes ● Other informal methods Assess modelling assumptions Formal analysis: ● Confidence intervals ○ Prediction intervals ● Significance tests ● Other formal methods	In the language of the Question context: Answer(s) Limitations Recommendations ['*Evidence*-based decisions, improvements, ...' means using Answers from *data*-based investigating with an *adequate* Plan *properly* executed.]
Output	***clear Question(s)***	***a Plan***	***Data***	***Information***	***Knowledge***

STATISTICS and STATISTICAL METHODS: Populations and Processes and ... (continued 1)

4. Populations and Processes: A Broader Perspective

As discussed on pages HL94.1 and HL94.2 of this Highlight #94, in the *statistical* framework provided by the FDEAC cycle, a population is a group of elements that is the focus of the Question to be answered; distinguishing four populations by the adjectives target, study, respondent and non-respondent is a way to identify and then manage sources of error that impose limitations on Answers. Some statistical Questions are better answered by starting with a process rather than a population.

Considered more broadly, the concepts of 'population' and 'process' can also provide a framework for developing terminology for *probability* and, in doing so, shed light on the troublesome idea of 'randomness' and on issues of variation and uncertainty.



As shown in the main schema at the right, for *probability* we start with the idea of a **process** that has two or more possible **outcomes** (but only one of which occurs when the process is executed); the set of *all* possible outcomes is the **sample space**, which is made up of **points** (although they could equally well be called outcomes as already defined). An **event** is any subset of these points, technically including the 'subset' consisting of *all* the points, so the sample space can be considered an event. Common notation for the sample space is S – *we* use a Roman S to avoid confusion with the population (data) standard deviation $\mathbf{S}$ and the random variable S representing the sample standard deviation s under equiprobable selecting.

- The smaller (left-hand) schema at the right shows corresponding *statistical* terms, illustrating the (unfortunate) practice of different disciplines using different terms for the same (or similar) concepts. [Confusingly, 'sample' in 'sample space' has little in common with a 'sample' in statistics.]

Only in the case of 'sample' and 'event' is there a marked difference in intent, despite their commonality of being subsets.

Often there is *repetition* of a process, with the recognition that which of the possible outcomes is obtained on any repetition is *un*predictable. Repetition of a process is sometimes called *repeating an experiment* – for example, see the definitions of the law of large numbers cited on page HL94.5. This useage should be avoided; 'experiment' should be kept *solely* for the vital distinction between experimental and observational Plans – see Statistical Highlight #89.

In introductory probability, the processes first discussed are often tossing a coin, rolling a die or selecting cards from a deck or balls from an urn; repetition of the process is often involved.

- A coin toss has two possible outcomes, a head (H) or a tail (T). When there is repetition of the tossing process, for *two* tosses there are four outcomes HH, HT, TH TT and, for three tosses, eight outcomes HHH, HHT, HTH, THH, HTT, THT, TTH, TTT. The event *at least two heads in three tosses* would comprise the first four of these eight points (or outcomes).
- The process of rolling a die once has six possible outcomes 1, 2, 3, 4, 5, 6; the event *an even outcome* would comprise the points 2, 4, 6. If the process is rolling the die twice (or rolling two dice once), there are $6 \times 6 = 36$ possible outcomes although only eleven possible dice totals 2, 3, ..., 12. [Which of these two we are concerned with depends on context.]
- In the version of poker where each player is dealt a 5-card hand, there are $\binom{52}{5} = 2{,}598{,}960$ possible hands ('outcomes'); this number of *minutes* is just under 5 years so an avid poker player might, in a lifetime, see all possible 5-card poker hands.
- The number of possible bridge hands is $\binom{52}{13} = 635{,}013{,}559{,}600$; this number of *seconds* is a little over 20,000 years. The large number of possible hands is one factor that makes bridge an intellectual challenge.

We discuss below selecting balls from an urn; 'urn' has traditionally been used in this context for a container of balls.

The final step in this introduction to the terminology of probability is to assign a *probability* – a number or symbol like $f(y_j)$ – to each point of the sample space, subject to the constraint that these probabilities sum to 1 over all the points of the sample space; this constraint is expressed symbolically as Pr(S) = 1. Illustrations of some probabilities are as follows.

∗ For a so-called *fair* coin, Pr(H) = Pr(T) = ½.
  For two and three tosses of a fair coin, the four and eight outcome listed above each have, respectively, a probability of ¼ and ⅛.
  The probability of the event *at least two heads in three tosses* is the sum of four probabilities of ⅛ which is ½.
  – It is useful to remember that, while ordinary coins are usually close enough to fair for many practical purposes, an *exactly* fair coin is an idealization, likely *un*attainable in practice; the same is true for a fair die.

∗ For one roll of a fair die, each outcome has probability ⅙.
  For two rolls of a fair die, each of the 36 possible outcomes has probability $1/36$.
  The eleven possible totals have respective probabilities $1/36$, $2/36$, $3/36$, $4/36$, $5/36$, $6/36$, $5/36$, $4/36$, $3/36$, $2/36$, $1/36$; thinking of each total as an event, the numerators of these probabilities are the number of ways the total can arise and so it is the sum of the probabilities of the relevant number of points each with probability $1/36$.

5. Selecting Equiprobably from an Urn Model

Suppose there is an urn which contains three red and three black balls; two balls are selected equiprobably with*out* replaaement. The probability of getting either two red or two black balls is $\binom{3}{2}\binom{3}{0}/\binom{6}{2} = 3/15$;
the probability of getting one ball of each colour is $\binom{3}{1}\binom{3}{1}/\binom{6}{2} = 9/15$.

(*continued overleaf*)

At first sight, it may seem surprising that the composition of an *individual* two-ball selection is unpredictable, but the *long-term* proportions of the three possibile compositions, as given by the three probabilities, *are* predictable.

To understand this, we first distinguish the three red and the three black balls in the urn by, say, thinking of each set having labels 1, 2 and 3; there are then 15 possible two-ball samples: R_1R_2, R_1R_3, R_2R_3, B_1B_2, B_1B_3, B_2B_3, R_1B_1, R_1B_2, R_1B_3, R_2B_1, R_2B_2, R_2B_3, R_3B_1, R_3B_2, R_3B_3. We see immediately that the proportions of the three sample compositions are, as realized under *equi*probable selecting, the three probabilities calculated above. This shows that if we were to replace the original urn with one containing 15 dumbells, three red (representing a sample with two red balls), three black (representing a sample with two black balls) and nine half red and half black (representing a sample with a red and a black ball), the probability structure is *identical* to that of the original urn with six balls, under the continuing requirement for *equi*probable selecting.

Thus, the set (really, the *population*) of all possible samples demonstrates that:

+ *un*predictable *individual* outcomes arise from *equi*probable selecting a *sample* (a *sub*set) in the presence of *variation* among population elements (in this instance, variation in ball colour),
+ the predictable *long-term* behaviour mirrors, under equiprobable selecting, the structure of the set of all possible samples.

These insights provided by invoking, in the context of an urn model and selecting with*out* replacement, the set of all possible samples should help us recognize (and avoid) unhelpful statements like:

– *A* random phenomenon *has outcomes that we cannot predict but that have a regular distribution in very many repetitions. This remarkble fact is the basis for the idea of probability.*

– *A random event, by definition, does not lend itself to explanation, but collections of random events do behave in a highly regular fashion.*

NOTES: 3. The foregoing insight from the set of all possible samples is like that from the Fundamental Theorem of Statistics in Appendix 3 on page HL74.5 in Statistical Highlight #74.

4. For *our* urn model, it happens that the probability structure is also retained if the six-ball urn is replaced by one with *five* dumbells – one red, one black, three mixed – but the demonstration is clearer with 15 dumbells.

5. The second unhelpful statement given above may (unintentionally?) imply that 'randomness' is somehow inimical to causation. That this is a dangerous misconception is illustrated in *The Biology of Cancer. A New Approach* by P. R. J. Burch. 452 pages (MTP, Lancaster, 1976). The book argues that the major diseases of affluent societies are caused by chance events scattered 'randomly' among what is postulated to be the body's central controlling cells (perhaps lymphocytes) and so these events cannot be due to external agents. *If* this were the case, a consequence is that the current major causes of morbidity and death cannot be treated by lifestye or pharmacological interventions. Two excerpts from a review of this book by John Cairns, *Riding roughshod over cancr research. Book review supplement: Nature* **200**, p. 198, March 18, 1976, are noteworthy:

> Now, all of this curve fitting would be just innocent fun, albeit of a rather old-fashioned kind, were it not that somehow the idea of chance events scattered randomly among central controlling elements has been taken to mean that the events cannot be due to external agents. I must say I do not understand the logic here; after all, one of the first collections of biological data shown to follow the Poisson distribution was the number of men killed by horses each year in the Prussian army, and there the external agent was plain for all to see, on the hoof as it were. ....
>
> Now, I believe that all this is what might euphemistically be called wrong. ... Were the biology of cancer a normal branch of science, (Burch's) unorthodox beliefs might be stimulating and would add a certain piquancy to the lives of cancer research workers, ... Unfortunately, the biology of cancer is not that kind of subject. This book will undoubtedly persuade many people to continue smoking who might otherwise have stopped, and may confuse those who legislate on the search for and control of carcinogens in our environment. Some of these people will, I am sure, have been persuaded not by actually reading the book and being swayed by the inexorable logic of his argument, but simply by the thought that when a professor writes a fat, obscurely mathematical tome on any subject whatsoever, there must be something in what he is saying. And that is why I wish this book had not been written.

6. The discussion overleaf on page HL94.3 involves the set of 15 *un*ordered possible samples from the six-ball urn; *ordered* discussion would involve 30 samples – for instance, we would distinguish R_1R_2 from R_2R_1 – but the probability structure would be unaffected. The ordered samples are relevant for comparison to the discussion below when selecting is *with* replacement.

- Although the discussion below raises matters of statistical interest, selecting with replacement is usually invoked only to gain mathematical tractability and, because the same unit can potentially be selected more than once for the sample, it is rarely used in statistical contexts – but see Note 4 on page HL38.7 in Statistical Highlight #38.

Consider again the six-ball urn containing three red and three black balls with two balls being selected equiprobably but now *with* replacement – that is, any ball selected is put back in the urn before the next selection is made; as a consequence, the urn *always* contains three red and three black balls so the selecting could, in principle, go on indefinitely instead of having to stop when the urn is empty as can happen when the selecting is with*out* replacement.

With equal numbers of red and black balls always in the urn, the probability of selecting a red or a black ball is always ½, so

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Populations and Processes and ... (continued 2)

the probability of selecting two red balls or of selecting two black balls is ½×½ = ¼, and the probability of selecting one red and one black ball is ¼ + ¼ = ½, corresponding to selecting a red ball followed by a black or a black ball followed by a red. These probabilities of ¼, ¼ and ½ for the three possible sample compositions are *different* from those – ⅕, ⅕ and ⅗ – from the previous selecting with*out* replacement (on page HL94.3).

Although the three probabilities differ in value when selecting without and with replacement, we see again in the latter case unpredictable individual outcomes but long-term regularity which reflects an underlying structure. These two matters are the start of similarities and some (surprising) differences arising from what might seem to be a minor change in the selecting process.

- The probability structure is retained for an urn containing *any* (even non-zero) number of balls, provided the number of red balls and black balls is the same; it is easiest for our purposes to have only two balls, one red and one black. The four possible samples – RR, BB, RB, BR – have the same proportions, under equiprobable selecting, as the three probabilities.

  Thus, as for selecting with*out* replacement, the set (the *population*) of all possible samples demonstrates that:

  + *un*predictable *individual* outcomes arise from *equi*probable selecting a *sample* in the presence of *variation* among population elements (variation in ball colour),
  + the predictable *long-term* behaviour mirrors, under equiprobable selecting, the structure of the set of all possible samples.
- With no particular advantage in *this* case, we could also consider an urn containing four dumbells – one red, one black, one red and black but more red, another red and black but more black, the latter two representing the *ordered* selecting of the mixed composition samples RB and BR.
- The two-ball urn shows that any urn with equal numbers of red and black balls and selecting *with* replacement is probabilistically equivalent to tossing a fair coin; selecting two balls is equivalent to two tosses, whose four outcomes (a 4-point sample space) HH, HT, TH, TT were given before near the middle of page HL94.3. Selecting with replacement from such an urn model is thus another instance of Bernoulli trials, defined on page HL4.1 and illustrated as three binomial distribution probability functions on page HL4.3 in Statistical Highlight #4.
- The long-term regularity in the case of repetition of a process with two equiprobable outcomes (say 'success' and 'failure') has one (surprising) feature – as the *proportion* of each outcome tends to become progressively closer to ½ as the number of repetitions increases, their *deviation* from half the number of repetitions tends to *in*crease; an illustration is as follows:

Number of repetitions	10	100	1,000	10,000	100,000	1,000,000
Number of successes	6	55	530	5,200	51,500	505,000
Number oa failures	4	45	470	4,800	48,500	495,000
Proportion of successes	0.6	0.55	0.53	0.52	0.515	0.505
Proportion of failures	0.5	0.45	0.47	0.48	0.485	0.495
Deviation from repetitions/2	1	5	30	200	1,500	5,000

  The *regular* trend of the proportion towards ½ in this illustration is only to make the main point clear; in any actual instance, the tend is usually (much) less regular.
- The behaviour of process outcomes under many repetitions is the context of **the law of large numbers**. A Google search in May, 2022, of this title yields a variety of results; for example:
  - ○ *As a sample size grows, its mean gets closer to the average of the whole population.* (Investopedia)
  - ○ *A theorem that describes the result of performing the same experiment a large number of times.* (Wikiedia)
  - ○ *If you repeat an expriment a large number of times, what you obtain should be close to the expected value.* (Probability Course)
  - ○ *As the number of identically distributed randomly generated variables increases, their sample mean (average) approaches their theoretical mean.* (Encyclopedia Britannica)
  - ○ *A theorem that describes the result of repeating the same experiment a large number of times.* (Corporate Finance)
  - ○ *If we take a sample of n observations of a random variable, the sample mean approaches the expected value of the random varable as n approaches infinity.* (Khan Academy)
  - ○ *The frequencies of events with the same likelihood of occurrence even out, given enough trials or instances. As the number of experiments increases, the actual ratio of outcomesa will converge on the expected, or theoretical, ratio of outcomes.* (Whatis.com)
  - ○ *As a sample size becomes larger, the sample mean gets closer to the expected value.* (Statology)
  - ○ *When we do an experiment a large number of times, the average result will be very close to the expected result. In other words, in the long run, random events tend to average out at the expected value.* (Math is Fun)
  - ○ *The relative frequency of an event will converge on the probability of the event, as the number of trials* increases. (Statistics Dictionary)

  It is of concern that a 'law' in probability has spawned such a variety of (often opaque and careless) 'definitions'.

  However, our two previous statements that predictable *long-term* behaviour mirrors, under equiprobable selecting, the structure of the set of all possible samples, expresses the same (general) idea, recognizing that the set of all possible samples comprises (exactly) multiple copies of the population elements and so accommodates the case of a *quantitative* sample and population attribute (like an average) as well as a categorical attribute (like ball colour in our urn model).

(*continued overleaf*)

There is also the 'law' (or 'myth') of *small* numbers – a facetious but useful reminder of the severe limitation on an Answer from a sample that is too small in the investigation context – but see Section 5 on page HL2.12 in Statistical Highlight #2.

- Selecting with replacement involves the idea of a process that can continue indefinitely; probability distributions, like the negative binomial and Poisson distributions illustrated in Statistical Highlight #4, whose random variable values can increase indefinitely (although with associated probabilites that are so small as to seldom be of practical interest), involve a similar idea. Such situations, perhaps in conjunction with the routine encounter with infinity in mathematics, may be responsible for the statistically *un*helpful concept of an infinite 'population'. This idea is 'normalized' by a term like the *finite population correction* in survey sampling where, of all places, the reality of populations as *finite* statistical entities should be obvious [see Note 2 on page HL77.5 in Statistical Highlight #77 and the discussion of equation (HL94.36) on page HL94.11].
  - It can help to avoid muddled thinking if a population (at least in statistics) is *only* a finite entity and we use *infnite population **model*** for an entity with no attainable end – see also the definition of a population early in Section 2 on page HL94.1
- Selecting *with* replacement – as in the discussion overleaf on page HL94.5 – raises the (technical) idea of **probabilistic independence**; the following matters are noteworthy in the context of this Highlight #94.
  - Its main usefulness is to make the probability of an intersection of probabilstically independent events the *product* of their individual probabilities; *we* used this property near the top of page HL94.5 to find the three probabilities of ¼, ¼ nd ½.
  - Probabilistic independence is a staple of introductory probability courses but usually with little emphasis of two matters.
    - It is a mathematical *idealization* – in the real world, *de*pendence is the (almost) universal experience;
    - It is usually troublesome to assess in the real world how well a modelling assumption of probabilistic independence is met.

6. Quantifying Variation for Data and for Random Variables

Discussion of the urn model with red and black balls in Section 5 on pages HL94.4 to HL94.6 shows that, whether selecting is without or with replacement, *one* of the factors responsible for unpredictable individual outcomes is *variation* among (response) variate values in the population elements (or units).

The measure of variation common in statistics is the *standard deviation*. For the respomse variate values in a sample, y_j ($j = 1, 2, ..., n$), and in a population, $\mathbf{Y}_i$ ($i = 1, 2, ..., \mathbf{N}$), the (data) s.d.s are as given in equations (HL94.1) and (HL94.2) at the right.

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2} \qquad \text{-----(HL94.1)}$$

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \bar{\mathbf{Y}})^2} \qquad \text{-----(HL94.2)}$$

A less familiar measure of variation is the *standard pairwise difference* (abbreviated s.p.d.) which is discussed in detail in Statistical Highlight #100. The (sample) s.p.d. can be based on the *un*ordered pairwise difference as in equation (HL94.3) or on the (twice as numerous) *ordered* pairwise differenes as in equation (HL94.4). The corresponding population expressions are equations (HL94.5) and (HL94.6). The *un*ordered expressions (HL94.3) and (HL94.5) are more convenient for the discussion in Statistical Highlight #100 but the ordered expressions will be used in this Highlight #94.

$$s_{pd} = \sqrt{\frac{2}{n(n-1)}\sum_{j<k=1}^{n}(y_j - y_k)^2} \qquad \text{-----(HL94.3)}$$

$$s_{pd} = \sqrt{\frac{1}{n(n-1)}\sum_{j\neq k=1}^{n}(y_j - y_k)^2} \qquad \text{-----(HL94.4)}$$

$$\mathbf{S}_{pd} = \sqrt{\frac{2}{\mathbf{N}(\mathbf{N}-1)}\sum_{i<l=1}^{\mathbf{N}}(\mathbf{Y}_i - \mathbf{Y}_l)^2} \qquad \text{-----(HL94.5)}$$

$$\mathbf{S}_{pd} = \sqrt{\frac{1}{\mathbf{N}(\mathbf{N}-1)}\sum_{i\neq l=1}^{\mathbf{N}}(\mathbf{Y}_i - \mathbf{Y}_l)^2} \qquad \text{-----(HL94.6)}$$

It is shown on pages HL100.1 to HL100.3 in Section 3 of Statistical Highlight #100 that the s.d. and the s.p.d. are *algebraically* equivalent – that is, they both quantify variation on the same basis (*un*like the interquartile range, for instance) – but they differ in *magnitude* by a factor of $\sqrt{2}$ as shown in equation (HL94.7) and by a factor of 2 in their squares [equation (HL94.8)]; that is, the s.p.d. is around 40% larger.

$$\text{s.p.d.} = \sqrt{2} \times \text{s.d.} \qquad \text{-----(HL94.7)}$$

$$\text{s.p.d.}^2 = 2 \times \text{s.d.}^2 \qquad \text{-----(HL94.8)}$$

$$\sum_{j=1}^{n}(y_j - \bar{y}) \equiv 0 \qquad \text{-----(HL94.9)}$$

The s.p.d. is useful here for three reasons.

+ Using the data directly as pairwise differences, rather than first finding their average and then using their deviations from it, the s.p.d. is a more 'natural' measure of variation than is the standard deviation.
+ *Un*like the s.d., its divisor is the number of entities being averaged which, taken with the algebraic equivalence of the two measures, justifies the divisor for the s.d. that takes account of there being one *constraint* on the deviations from the average – they add to zero regardless of the y values [equation (HL94.9)]. There is *no* similar constraint on the pairwise differences.
+ The s.p.d. using the ordered pairwise differences can be extended to include the n (or $\mathbf{N}$) *self*-differences, which are all (obviously) zero and so add nothing to the numerator; however, the divisor becomes n^2 (or $\mathbf{N}^2$) – see equations (HL94.10) and (HL94.11), where the subscript 'a' added to the symbol indicates an s.p.d. based on *all* pairwise differences. Comparing equations (HL94.10) and (HL94.11) with (HL94.4) and (HL94.6) shows that the relationships between the s.p.d. with and without the self-differences are equations (HL94.12) and (HL94.13).

$$s_{pda} = \sqrt{\frac{1}{n^2}\sum_{j,k=1}^{n}(y_j - y_k)^2} \qquad \text{-----(HL94.10)}$$

$$\mathbf{S}_{pda} = \sqrt{\frac{1}{\mathbf{N}^2}\sum_{i,l=1}^{\mathbf{N}}(\mathbf{Y}_i - \mathbf{Y}_l)^2} \qquad \text{-----(HL94.11)}$$

$$s_{pda} = \sqrt{\frac{n-1}{n}}\, s_{pd} \qquad \text{-----(HL94.12)}$$

$$\mathbf{S}_{pda} = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\, \mathbf{S}_{pd} \qquad \text{-----(HL94.13)}$$

The statistically bizarre measures of variation of equations (HL94.10) and (HL94.11) are equivalent to an s,d. with a divisor of n (or $\mathbf{N}$) and so make even *less* appealing the (misguided) idea of these divisors for a (data) s.d. – see Appendix 3 on pages HL100.7 and HL100.8 in Statistical Highlight #100. (A supposed rationale for this idea is to preserve the simple grade-school definition of averaging.)

  - Self-differences are reminiscent of selecting *with* replacement, adding to the oddity of equations (HL94.10) and (HL94.11).

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Populations and Processes and ... (continued 3)

Having dealt with *data* variation, we turn to **random variables**, which are a component of the probability *models* used in statistics; our concern here is primarily with *discrete* random variables which take on a *finite* number of values (which are often numerical but may be categorical). *We* denote random variables by upper-case *italic* letters, usually from near the end of the alphabet; we try to keep upper-case italic letters near the start of the alphabet for *events.*

- Formally, a random variable is an entity that assigns a probability to each point of the sample space.
- Informally, a random variable is an entity that takes on different values according to chance.

y	RR	RB	BB
$f(y)$	3⁄15;	9⁄15	3⁄15

;

z	RR	RB	BB
$f(z)$	¼	½	¼

Both statements mean that we can specify a random variable by a table that gives its values and their probabilities. For example, from the discussion in Section 5 on pages HL94.3 to HL94.6, the random variables Y and Z representing the probabilities of the three two-ball sample compositions from the six-ball urn selected without and with replacement are as shown at the right. The notation in these tables is a lower case *italic* letter (y, z) for the random variable *values* and f for the **probability function** that gives the *probabilities.* Other probability functions (p.f.s) are given in Statistical Highlight #4; note that, for the hypergeometric p.f.s on page HL4.1, the probability is *zero* for two random variable values.

Our use of f (rather than p) for a probability function allows us to keep the letter p for a *proportion*, with four meanings as shown at the right in Table HL94.3. The distinctions we can maintain with this notation are easily lost when the predominant use of letter p is P for probability, as is the usual practice in introductory statistics texts. In particular, our distinguishing p and p reminds us of the statistician's responsibility to develop a Plan for an investigation which makes it reasonable to treat sample proportions as values of a (model) random variable P.

Table HL94.3

QUANTITY	SYMBOL
Population proportion	Ᵽ
Sample proportion	p
Random variable	P
A value of P	p
Model parameter	π
Probability	Pr

As an entitty useful in probability modelling, a random variable has a mean and a standard deviation. For a random variable Y with k values – perhaps representing a response variate y or Ɏ so that k may be N̶ or n or some combination of them – the *mean* is as given in equation (HL94.14) at the right above. A mean is the same *idea* as an average but they are *not* identical, which is why *we* use *different* words for them.

$$\text{mean of } Y = \sum_{\mathrm{k}} y \bullet \Pr(Y = y) \qquad \text{-----(HL94.14)}$$

$$\text{average of } y\text{s} = \sum_{\mathrm{k}} y \bullet \frac{1}{\mathrm{k}} \qquad \text{-----(HL94.15)}$$

- The **mean** of a (discrete) random variable, Y say, is the sum of the products of its values and their probabilities [equation (HL94.16)]; thus, a mean is the *weighted* average of the random variable's values with their probabilities as the weights. The subscript Y on μ is needed only to avoid confusion when two or more random variables are involved.

$$\mu_Y \equiv E(Y) = \sum_{j=1}^{\mathrm{k}} y_j \times f(y_j) \qquad \text{-----(HL94.16)}$$

$$\sum_{j=1}^{\mathrm{k}} (y_j - \mu) \times f(y_j) = 0 \qquad \text{-----(HL94.17)}$$

  - Alternative terminology for the mean of Y is its **expected value**, denoted $E(Y)$, sometimes shortened to **expectation**.
  - If we subtract and add k terms $\mu \times f(y_j)$ on the right-hand side of equation HL94.16, we obtain:
    $\mu = \sum_{j=1}^{\mathrm{k}} (y_j - \mu) \times f(y_j) + \mu \times \sum_{j=1}^{\mathrm{k}} f(y_j) = \sum_{j=1}^{\mathrm{k}} (y_j - \mu) \times f(y_j) + \mu$ because $\sum_{j=1}^{\mathrm{k}} f(y_j) = 1.$ which yields equation (HL94.17) above.
  - Comparing equation (HL94.17) with equation (HL94.9) on the facing page HL94.6, we see that, while the sum of deviations from the *average*, $\sum(\mathrm{y_j} - \bar{\mathrm{y}})$, *is* zero, this general constraint holds only for the *weighted* differences of the values of a *random variable* from its *mean*; the *un*weighted sum $\sum(y_j - \mu)$ is *not* constrained to be zero. An alternative way of writing equation equation (HL94.17) as an *expected value* is equation (HL94.18); we must be careful not to confuse this *expectation* with the *sum* $\sum(y_j - \mu)$ or to mistakenly interpret equation (HL94.18) as the *same* constraint as equation (HL94.9) – equation (HL94.17) makes the distinction clearer.

$$E(Y - \mu) = 0 \qquad \text{-----(HL94.18)}$$

  - A special case when $\sum(y_j - \mu)$ *is* zero is when Y has a *uniform* distribution – its probabilities are all 1/k; the mean of Y is then the *average* of its values so the constraint (HL94.9) *does* apply.
    Random variables with *equi*probable outcomes arise again in discussion overleaf on page HL94.8.

- The **standard deviation** of a (discrete) random variable, Y say, is the sum of the products of the squared deviations of its values from its mean and the probabilities of these values [equation (HL94.19)]; thus, a (probabilistic) standard deviation of a random variable is the *weighted* average of its values' squared deviations from its mean with their probabilities as the weights.

$$\sigma_Y \equiv s.d.(Y) = \sqrt{\sum_{j=1}^{\mathrm{k}} (y_j - \mu)^2 \times f(y_j)} \qquad \text{-----(HL94.19)}$$

$$\sigma = \sqrt{E(Y - \mu)^2} \qquad \text{-----(HL94.20)}$$

$$\sigma = \sqrt{E(Y^2) - [E(Y)]^2} \qquad \text{-----(HL94.21)}$$

$$E(Y_1 - Y_2)^2 = 2\sigma^2 \qquad \text{-----(HL94.22)}$$

  - Equation (HL94.19)] can also be wrtten as an expected value [equation (HL94.20)]; this is usually wrtten as the more convenient algebraically equivalent equation (HL94.21).
  - Like means and averages, what *we* distinguish as probabilistic and data standard deviations have similar intent but differences in implementation, so it is *un*surprising if (usuallu minor) differences between them arise when their paths cross.
  - If random variables Y_1 and Y_2 both have mean μ and s.d. σ and are probabilistically independent:
    $E(Y_1 - Y_2)^2 = E(Y_1^2 - 2Y_1Y_2 + Y_2^2) = E(Y_1^2) - E(Y_1)E(Y_2) + E(Y_2^2) - E(Y_1)E(Y_2) = E(Y_1^2) - [E(Y_1)]^2 + E(Y_2^2) - [E(Y_2)]^2 = \sigma^2 + \sigma^2.$
    This is equation (HL94.22) above, which is reminiscent of the numerical relationships between the values of the s.p.d. and a (data) s.d. and their squares in equations (HL94.7) and (HL94.8), on the facing page HL94.6.

(*continued overleaf*)

The foregoing discussion of means and probabilistic standard deviations of random variables, and their comparison to averages and data standard deviations, is illustrated in two tables at the right below.

Table HL94.4 shows a *random variable* with k = 4 values and corresponding probabilities; its four values are the *data* for second line of table entries. We see that the mean of the random variable and the average of the data differ in value, as do the corresponding (probabilistic) and (data) standard deviations.

Table HL94.4: Random variable/Data Comparison

						..... Location			.... Scale (Variation)		
						Name	**Symbol**	**Value**	**Name**	**Symbol**	**Value**
Random variable	y_j	0	1	2	3	Mean	μ, $E(Y)$	1.3	S.d.	σ, $s.d.(Y)$	0.9
	$f(y_j)$	0.2	0.4	0.3	0.1						
(Population or sample) Data	$\mathbf{Y}_i$ or y_j = 0, 1, 2, 3					Average	$\overline{\mathbf{Y}}$, $\bar{y}$	1.5	S.d.	$\mathbf{S}$, s	$\sqrt{5/3} \simeq 1.3$

The change in Table Hl94.5 is that the four probabilities for the random variable are all *equal*; the mean and average are now equal and, although the two standard deviations still differ in value, their numerators are the same and only their divisors (k and k − 1) differ.

Table HL94.5: Random variable/Data Comparison

						..... Location			.... Scale (Variation)		
						Name	**Symbol**	**Value**	**Name**	**Symbol**	**Value**
Random variable	y_j	0	1	2	3	Mean	μ, $E(Y)$	1.5	S.d.	σ, $s.d.(Y)$	$\sqrt{5/4} \simeq 1.1$
	$f(y_j)$	0.25	0.25	0.25	0.25						
(Population or sample) Data	$\mathbf{Y}_i$ or y_j = 0, 1, 2, 3					Average	$\overline{\mathbf{Y}}$, $\bar{y}$	1.5	S.d.	$\mathbf{S}$, s	$\sqrt{5/3} \simeq 1.3$

- The (probabilistic) s.d. in Table HL94.5 is *larger* in magnitude than in Table HL94.4 because the extreme values of 0 and 3 of the random variable are more heavily weighted by their higher probabilities.
- Illustrating the discussion around equation (HL94.18) overleaf on the lower half of page HL94.7, we see that $\Sigma(y_j - \mu)$ is 0.8 in Table HL94.4 and 0 (for the equiprobable case) in Table HL94.5.

7. The Average of Random Variables

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + ... + Y_n) \qquad \text{-----(HL94.23)}$$

$$E(\overline{Y}) = \mu \qquad \text{-----(HL94.24)}$$

$$s.d.(\overline{Y}) = \sigma/\sqrt{n} \equiv \sigma\sqrt{\frac{1}{n}} \qquad \text{-----(HL94.25)}$$

Suppose we have a random variable $\overline{Y}$ that is the *average* of n probabilistically independent random variables Y_1, Y_2, ..., Y_n; each of these n random variables has mean μ and standard deviation σ [equation (HL94.23)]. The mean and standard deviation of $\overline{Y}$ are μ (the *same* as its n components) and $\sigma/\sqrt{n}$ (smaller – *less* variable – than its components but only by a factor of the *square root* of their number). [The derivations of these two results are given on page 5.36 in Figure 5.14 of the STAT 220 Course Materials.]

- The choice of n for the number of random variables in the average (HL94.23) is because, ultimately (in Section 8 below), equations (HL94.23) to (HL94.25) will be discussed in a *(survey) sampling* context.
- It is easy for forget that these two (model) results (HL94.24) and (HL94.25) are for random variables with the (*un*worldly) assumption of probabilistic independence.
- As we will discover in Section 8 below, randon variables have (also unworldly) overtones of selecting *with* replacement.

8. The Distribution of the Sample Average under Equiprobable Selecting

Random variables (as models) arise naturally in the context of *probability selecting* of a samplr of units from a population; here, we confine our attention to *equiprobable* selecting. Statistical Highlight #77 presents survey sampling theory for estimating a population average $\overline{\mathbf{Y}}$ or total ${}_T\mathbf{Y}$; our interest here is with results from pages HL77.4 and HL77.5.

Our first concern is with **means**. We define two random variables:

∗ Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over; -----(HL94.26)

∗ $\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + ... + Y_n)$, which is equation (HL94.23); $\overline{Y}$ represents the distribution of the sample average under equiprobable selecting (with*out* replacement).

The definition of Y_j reminds us that *measuring* is (always) involved in generating data but our focus here is on *selecting*; however, measuring can be modelled as equiprobable selecting – see Section 5 on page HL74.3 of Statistical Highligh #74.

$$\Pr(Y_j = \mathbf{Y}_i) = \frac{1}{\mathbf{N}} \qquad \text{-----(HL77.4) = (HL94.27)}$$

$$E(Y_j) = \overline{\mathbf{Y}} \qquad \text{-----(HL77.5) = (HL94.28)}$$

$$E(\overline{Y}) = \overline{\mathbf{Y}} \qquad \text{-----(HL77.8) = (HL94.29)}$$

Three results from Statistical Highlight #77 are equations (HL94.27) to (HL94.29) at the right. we see that Y_j has $\mathbf{N}$ equiprobable values [the population element (or unit) responses], and it and $\overline{Y}$ both have mean $\overline{\mathbf{Y}}$. [The last of these results is the unbiassedness of $\overline{Y}$ as an etimator of $\overline{\mathbf{Y}}$].

- The mean and the average being *equal* here is a reminder of Table HL94.5 above for the case of a random variable with equiprobable values.

Interestingly, these results are *un*changed if the equiprobable selecting is *with* replacement; we see why this is so from

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Populations and Processes and ... (continued 4)

Tables HL94.6 and HL94.7 at the right and below, which list the set of all possible ordered samples of size 2 and size 3 from a population of four elements (or units), when the equiprobable selecting is without (on the left side of each table) and with (on the right side) replacement. The respective numbers of samples are 12 and 16 in Table HL94.6 and 24 and 64 in Table HL94.7.

Table HL94.6: Set of All Possible Samples;
$\mathbf{N} = 4$, $\mathbf{Y}_i = 1, 2, 3, 4$; **Sample size n** $= 2$

Without replacement						**With replacement**									
(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)	(1,1)	(2,2)	(3,3)	(4,4)
(2,1)	(3,1)	(4,1)	(3,2)	(4,2)	(4,3)	(2,1)	(3,1)	(4,1)	(3,2)	(4,2)	(4,3)				

Table HL94.7: Set of All Possible Samples; $\mathbf{N} = 4$, $\mathbf{Y}_i = 1, 2, 3, 4$; **Sample size n** $= 3$

Without replacement						**With replacement**																		
(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,1,2)	(3,2,1)	(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,1,2)	(3,2,1)	(1,1,2)	(1,1,3)	(1,1,4)	(1,2,2)	(2,2,3)	(2,2,4)	(1,3,3)	(2,3,3)	(3,3,4)	(1,4,4)	(2,4,4)	(3,4,4)	(1,1,1)
(1,2,4)	(1,4,2)	(2,1,4)	(2,4,1)	(4,1,2)	(4,2,1)	(1,2,4)	(1,4,2)	(2,1,4)	(2,4,1)	(4,1,2)	(4,2,1)	(1,2,1)	(1,3,1)	(1,4,1)	(2,1,2)	(2,3,2)	(2,4,2)	(3,1,3)	(3,2,3)	(3,4,3)	(4,1,4)	(4,2,4)	(4,3,4)	(2,2,2)
(1,3,4)	(1,4,3)	(3,1,4)	(3,4,1)	(4,1,3)	(4,3,1)	(1,3,4)	(1,4,3)	(3,1,4)	(3,4,1)	(4,1,3)	(4,3,1)	(2,1,1)	(3,1,1)	(4,1,1)	(2,2,1)	(3,2,2)	(4,2,2)	(3,3,1)	(3,3,2)	(4,3,3)	(4,4,1)	(4,4,2)	(4,4,3)	(3,3,3)
(2,3,4)	(2,4,3)	(3,2,4)	(3,4,2)	(4,2,3)	(4,3,2)	(2,3,4)	(2,4,3)	(3,2,4)	(3,4,2)	(4,2,3)	(4,3,2)													(4,4,4)

We see that, regardless of sample size and whether the selecting is without or with replacement, the set of all possible samples consists of multiple copies of the population response variate values, and so the average of each set *is* $\overline{\mathbf{Y}}$ (*i.e.*, unbiased estimating of $\overline{\mathbf{Y}}$ by $\overline{Y}$). Like the discussion at the top of page HL94.4, this is another insight provided by the set of all possible samples.

- The number of population copies differs, of course, within and between tables, being 6 and 8 in Table HL94.6 and 18 and 48 in Table HL94.7.
- Table HL94.6 has different *samples* in its rows, different *orders* of elements within samples in the columns; Table HL94.7 is like this in the second block on its right side but, for convenience of table format, has the reverse in its other three blocks.

Our second concern is with **standard deviation** and the random variables Y_j and $\overline{Y}$ of equations (HL94.26) and (HL94.23) on the facing page HL94.8

Three results from Statistical Highlight #77 are equations (HL94.30) to (HL94.32) at the right; they raise the following matters.

$$s.d.(Y_j) = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S} \qquad \text{-----(HL77.6) = (HL94.30)}$$

$$cov(Y_j, Y_l) = -\frac{\mathbf{S}^2}{\mathbf{N}} \qquad \text{-----(HL77.7) = (HL94.31)}$$

$$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\text{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(HL77.9) = (HL94.32)}$$

$$s.d.(\overline{Y})_{\text{n}=1} = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2} \qquad \text{-----(HL94.33)}$$

- The (probabilistic) standard deviation of Y_j simplifies to what looks like a *data* standard deviation but with a divisor of $\mathbf{N}$ insead of the (correct) $\mathbf{N}-1$. [The *same* square root term occurs in equation (HL94.13) near the bottom of page HL94.6.]
- The (non-zero) covariance of Y_j and Y_l reflects the effect on later selecting of what has previously been selected when the selecting is with*out* replacement {a key difference from selecting *with* replacement, which produces probabilistic independence and so the covariance is zero [but see the last comment (**+**) on the lower half of page HL94.10]}.
- The (probabilistic) standard deviation of $\overline{Y}$, derived (in Statistical Highlight #77) by taking account of the unit inclusion probabilities generated by the (equiprobable) selecting process (the so-called 'design-based' approach), contains an additional term under the square root involving $\mathbf{N}$, compared to the (model-based) expression in equation (HL94.25) on the facing page HL94.8.
  - Equation (HL94.25) is a staple of introductory statistics courses because of its (valuable) quantitative insight about the variation of an average. In light of its background of the mathematical idealization of probabilistically independent random variables, this insight pales before the richer insights of equation (HL94.32) of the same *strong* dependence on n of the variation of $\overline{Y}$ and its *weak* dependence on $\mathbf{N}$ in the usual real-world situation of $\text{n} \ll \mathbf{N}$.
    - The statistical accolades usually accorded equation (HL94.25) seem mainly to reflect its good fortune that, in the real world, typically $\text{n} \ll \mathbf{N}$.
  - When $\text{n} = 1$, equation (HL94.32) can be wrtten as equation (HL94.33) with. like equation (HL94.30), what looks like a (wrong) data standard deviation on its right-hand side. Equation (HL94.33) has been interpreted as a reason to calculate the population (data) standard deviation with a denominator of $\mathbf{N}$ insead of (the correct) $\mathbf{N}-1$; this idea should be (immediately) dismissed because it:
    - ignores the (simple) average in equation (HL94.5) on page HL94.6 being algrebraically equivalent to the (constrained) average in equation (HL94.2);
    - disregards equations (HL94.11) and (HL94.13) which show that a divisor of $\mathbf{N}$ in a (data) standard deviation corresponds to the statistically bizarre inclusion of the pairwise *self*-differences when quantifying (data) variation;
    - mistakenly assumes that probabilistic and data standard deviations are (exactly) equivalent – recall the discussion on the upper half of the facing page HL94.8;
    - produces the anomaly of differing calculations for the population and sample (data) standard deviations;
    - is based on the special case of a (probabilistic) standard deviation of random variables with *equiprobable* values.

  There is further discussion of this erroneous idea on the lower half of page HL100.8 in Statistical Highlight #100.

Statisical Highlight #77, the source of equations (HL94.30) to (HL94.32), involves selecting with*out* replacement, the method of selecting appropriate in most real-world contexts. For *this* Highlight #94, selecing *with* replacement would yield, from equation

 (*continued overleaf*)

(HL77.9) with the covariance terms all zero, equation (HL94.34) at the right.

$$s.d.(\overline{Y}) = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}\sqrt{\frac{1}{\text{n}}} \quad \text{-----(HL94.34)}$$

- This equation is the design-based analogue of the model-based equation (HL94.25) on page HL94.8, where σ corresponds to the first two components on the right-hand side of equation (HL94.34).
  - The first square root on the right-hand side of equation (HL94.34) deals with the (usually slight) difference between probabilistic and data standard deviations (as discussed on pages HL94.7 and HL94.8).
    - In the usual real-world situation of large $\mathbf{N}$, this square root is (extremely) close to 1 in value, but lack of practical effect is *not* a reason to disregard it.
    - Involvement of $\mathbf{N}$ and $\mathbf{N}-1$ in this square root is a reminder of what we see for the two standard deviations in Table HL94.5 on page HL94.8 for a random variable with equiprobable values.
- The correspondence between probabilisic independence of the random variables that lead to equation (HL94.25) and equiprobable selecting *with* replacement that leads to equation (HL94.34) is easily overlooked; this is unfortunate when it is selecting with*out* replacement (which generates probabilistic *de*pendence) that is of interest in most data-based investigating.

9. Issues in Teaching Introductory Statistics

The discussion in this Highlight #94 raises several matters relevant to teaching introductory statistics.

∗ Averages are of wide interest in statistical methods and so their behaviour, particularly their variation in relation to that of their components, is important statistically. The random variable $\overline{Y}$ [equation (HL94.23)] is discussed in Section 7 on page HL94.8 and survey sampling theory for estimating $\overline{\mathbf{Y}}$ is summarized (from Statistical Highlight #77) in Section 8, starting on page HL94.8. This is the usual order of material in an introductory text:

- random variables as models are introduced as part of introductory probability (often in Chapter 4 of such texts);
- this leads on to using $\overline{Y}$ to introduce the distribution of the sample average with mean and standard deviation as in equations (HL94.24) and (HL94.25) near the middle of page HL94.8;
- equation (HL94.32) (overleaf in the middle of page HL94.9), for the (more informative design-based) standard deviation of $\overline{Y}$ is likely not given or, if it is, its derivation is probably omitted.

It might be better pedagogy to reverse the order of how $\overline{Y}$ is introduced by starting with survey sampling which will be familiar to many. Earlier discussion will (hopefully) have distinguished the population attributes $\overline{\mathbf{Y}}$ and $\mathbf{S}$ from the sample attributes $\overline{\text{y}}$ and s and so, invoking the idea of equiprobable selecting without replacement, estimating $\overline{\mathbf{Y}}$ can be discussed with the following advantages.

+ Random variables (with equiprobable values, which make means and averages equal) arise naturally from the equiprobable selecting process.
+ The discussion is grounded in the (real-world) attributes $\overline{\mathbf{Y}}$ and $\mathbf{S}$, $\overline{\text{y}}$ and s;
+ The equiprobable selecting process (a component of the Plan) makes it clear why it is reasonable to regard $\overline{\text{y}}$ (the average of the sample actually selected) as a value $\overline{y}$ of the random variable $\overline{Y}$;
+ Diagrams like those in Statistical Highlights #74 (pages HL74.6 and HL74.7) and #76 help make clearer the idea of a sampling distribution – this is the *third* encounter in this Highlight #94 with the (useful) idea of the set of all possible samples (recall discussion near the top of pages HL94.4 and HL94.9).
+ *Un*biased estimating could be illustrated with*out* involving random variables by (again) the set of all possible samples as multiple copies of the population – recall the comment overleaf on page HL94.9 below Table HL94.7.
+ The effect of $\mathbf{S}$, n and $\mathbf{N}$ on the variation of $\overline{Y}$ is shown by equation (HL94.32).
  - The *derivation* of equation (HL94.32) could be given as enrichment for interested students as an appendix in the (later) introduction to $\overline{Y}$, the average of random variables (as in Section 7 on page HL94.8).
+ With averages firmly established from discussion of survey sampling, it might be possible to make inroads into the routine malpractice of calling averages 'means', and to reserve the latter term for random variables.
+ Background in real-world survey sampling should foster a better appreciation of the maxim *all models are wrong, some are useful* when random variables are recognized as having means and (probabilistic) standard deviations that differ from averages and (data) standard deviations and, for the *average* of random variables, the (unappealing) selecting *with* replacement that is implied by routinely-invoked probabilistic independence [as discussed above in the context of equation (HL94.34)].

∗ Equation (HL94.23) (near the middle of page HL94.8) may be (carelessly) written (in our notation) as equation (HL94.35) at the right below, about which a (puzzled) student might reasonably ask: *How can a number (on the left-hand side) be equal to a combination of random variables?* A (terse) statement accompanying equation (HL94.35) like: *If the population has mean μ, then μ is the mean of each observation Y_j*, compounds the confusion in multiple ways:

$$\overline{\text{y}} = \frac{1}{\text{n}}(Y_1 + Y_2 + \dots + Y_{\text{n}}) \quad \text{-----(HL94.35)} \qquad \textbf{!}$$

- a random variable is not a 'population';
- an average should not be called a 'mean';
- the *model* parameter μ is not a (*real-world*) attribute, like an average;
- the random variable Y_j is not an 'observation';
- an observation would not typically be thought of as having a 'mean'.

A lesson from this accompanying statement may be that brevity, like carelessness, can be the enemy of clarity – see also the

STATISTICS and STATISTICAL METHODS: Populations and Processes and ... (continued 5)

comment in the fourth bullet (●) on the upper half of page HL91.4 in Statistical Highlight #91. It also shows how easily discussion of a random variable like $\overline{Y}$ can lose coherence when the terminology of survey sampling is carelessly invoked, in place of survey sampling using random variables for probability modelling (as in Statistical Highlight #77).

∗ Statisticians should guard against allowing the deductive certainty, and general aura of inevitability, of mathematical modelling in statistical theory to distract them from the real-world focus of statistical methods; a symptom of this is when the results in equations (HL94.24) and (HL94.25) are described as (unqualified) 'facts'. Statistics teaching often seems to inply that the mathematics is really what matters and the preliminaries of terminology and notation do not merit the attention they so clearly need if we are to preserve (and convey to students) the intellectual integrity of statistics as a (vital) discipline.

– An illustration is when what these Materials call *study error* – the difference, in a particular investigation, between the attribute(s) in interest in the target population and the study population – is called *selection bias*, despite having little to do with either selecting or bias, with the comment that some forms of selection bias can be reduced by probability selecting methods.

∗ An illustration of the distortion arising from undue regard for modelling is when equation (HL94.32) (near the middle of page HL94.9) is written as equation (HL94.36) at the right; f is the sampling fraction $\mathrm{n}/\mathbf{N}$ and $1-f$ is called the *finite population* ***correction***. This raises three issues.

$$s.d.(\overline{Y}) = \sqrt{(1-f)}\,\mathbf{S}\sqrt{\tfrac{1}{\mathrm{n}}} \quad \text{-----(HL77.13) = (HL94.36)}$$

$$s.d.(\overline{Y}) = \sqrt{\frac{\mathbf{N}-\mathrm{n}}{\mathbf{N}-1}}\sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}\sqrt{\tfrac{1}{\mathrm{n}}} \quad \text{-----(HL94.37)}$$

– The *model* equation (HL94.25) has been given primacy over the more informative (design-based) equation (HL94.32), which has been relegated to being a 'correction' to the model expression.

– Expressed another way, the (model) idea of an *infinite* 'population' has taken precedence over that of a real-world *finite* population – recall the comment (●) near the top of page HL94.6.

– Equations (HL94.36) and (HL94.25) are *not* strictly equivalent because $\mathbf{S}$ and σ are not equivalent – recall equation (HL94.30) near the middle of page HL94.9. Taking account of this, equation (HL94.36) becomes equation (HL94.37), where the 'correction' (the first term on the right-hand side) is less convenient in form.

∗ At least in introductory courses, the intellectual challenges of understanding statistical methods are currently, in large measure, those that arise from presentation inadequacies, most notably careless terminology and notational poverty (like the absence of $\overline{\mathbf{Y}}$). It would be more profitable to provide opportunity for students to engage with the *actual* intellectual issues of statistics by providing presentations free from carelessness and confusion, even though this may entail more lengthy discussion and demand more discipline from both instructors and students.

Three precepts should inform statistics teaching in its choice of topics, their order, their terminology and notation.

- Do not confuse the population with the sample [and remember populations are *finite*, infinite 'populations' are (sometimes useful) models].
- Do not confuse the model with the real world (*e.g.*, an infinite 'population' is *not* the real world).
- Do not confuse the individual case with behaviour under repetition (the latter arises once probabilities become involved).

Their usefulness is illustrated in discussion on pages HL100.7 and HL100.8 in Statistical Highlight #100. These precepts were brought to the writer's attention by his colleague Prof. R.J. Mackay, likely in the 1980s. Sadly, there is little evidence of their observance in current introductory teaching materials; rather, their violation seems commonplace.

These precepts are reminiscent of fundamentals in other disciplines, like the Postulates of Impotence in physics – see pages 5.83 and 5.84 in Figure 5.7 of the STAT 231 Course Materials.

∗ The variety of introductory statistics courses in other disciplines implies that the usefulness of statistical methods is widely recognized. However, disciplines other than statistics (like engineering and medicine) with serious practical implications would realize that practitioners with one (or perhaps two) courses in the discipline would be likely to generate answers that are unhelpful, possibly dangerously so (see Statistical Highlight #31). It is curious that statistics has allowed itself to become a discipline where one course is considered enough, at least by some, to produce competence.

10. Can We Account for Long-term *Poisson* Regularity?

In a (simple) urn model and selecting without and with replacement, we can account straight forwardly for unpredictable short-term and regular long-term behaviour. It would be satisfying if we could also account for why the Poisson model is a (surprisingly) good fit to data sets for three real-world phenomena. Relevant information is summarized in Table HL94.8 at the right; the three data sets are given in Figure 10.6 on pages 10.19 to 10.21 in the STAT 221 Course Materials.

Table HL94.8

Context: Process	Domain	Unit	Outcome: Number		Observations: Total	Average
Horse kicks	Time	Year	Death	122	200	122/200 ≃ 0.61
Radioactivity	Time	7½ seconds	Emission	10,094	2,608	10,094/2,608 ≃ 3.8704
VI buzzbombs	Area	¹⁄₁₆ km²	Hit	537	576	537/576 ≃ 0.932 232

- Horse kicks refers to the number of fatal horse kicks in ten corps of the Prussian cavalary over a 20-year period; this is the data set referred to in Note 5 in the first excerpt quoted from the book review discussed on page HL94.4.
- Radioactivity refers to Rutherford and Geiger's 1910 counts of α-particles emitted by polonium over 2,608 periods of one-

 (*continued overleaf*)

eighth of a minute (*i.e.*, 7½ seconds).

- V1 buzzbombs refers to V1 flying bomb hits in a 36 square-kilometre area of London during World War 2; for the post-war data collection, the area was considered in ¼×¼ km squares. (A V1 fell to earth when it ran out of fuel.)

∗ Unpredictable short-term behaviour is the actual number of fatal horse kicks in a corps in a particular year, the number of α-particles emitted in a particular 7½-second period, and the number of V1 hits in a particular ¼×¼ km square.
  - These can be explained by the haphazard coming together of conditions essential for an outcome; for instance:
    - a cavalaryman standing where he could be kicked and the horse being startled or kicking for another reason;
    - fortuitously appropriate timing among nuclear processes needed for α-particle emission from polonium;
    - particularities of the V1 launch ramp, the amount of fuel in the V1 and the atmospheric conditions during its flight;

∗ Regular long-term behaviour is each data set being a reasonable fit to a Poisson model, with the caveat that the value used for the Poisson mean μ is the average from the data – see the last column of Table HL94.8 overleaf near the bottom of page HL94.11; the exact *form* of the 'regularity' is (of course) influenced by this value (see page HL4.4 in Statistical Highlight 4).
  - As discussed near the top of page HL94.5, in the urn model the long-term regularity reflects the structure of the set of all possible samples. In the three real-world data sets, 'selecting' and 'samples' are not so easily recognizable; however, one possibility is that what corresponds to the set of all possible samples is the possible ways that relevant conditions can come haphazardly together to generate the outcome – a potentially large number of these could account for the unbounded upper limit of the random variable of the Poisson model. It seems plausible that, in data sets like the three under discussion, the relevant conditions would remain (more or less) the *same* over time so that their ways of haphazard coming together might represent a persistent 'structure'' that would be reflected in long-term regularity. Depending on the number of conditions, which determines in how many ways the relevent outcome can arise, fewer or more repetitions would be needed for the putative 'regularity' to be exhibited. Presumably, the greater the number of repetitions, the more likely we would be to see 'regularity' if this suggestion for its origin has merit.

Three further speculative comments are that:

+ if the previous suggestion, or others, are deemed unconvincing, we may ultimately be left with Poisson 'regularity' in real-world phenomena (like the three discussed here) as simply another instance of the surprising (almost miraculous) ability of mathematics to describe the real world;
+ observed frequencies of more than 1 in 'Poisson' data could be viewed as the selecting *with* replacement which [as discussed in the second bullet (●) near the top of page HL94.10] is a real-world process that is able to justify the probabilistic independence assumption involved in the derivation of the Poisson distribution – see the bottom of page HL4.1 and the upper half of page HL4.2 in Statistical Highlight #4;
+ Feller comments (opaquely) [see page 10.21 of Figure 10.6 in the STAT 221 Course Materials] that the fit of the Poisson model to the V1 data indicates 'perfect randomness' (a phrase reminiscent of 'very unique'), as opposed to the contemporary view of *clustering* of hits in squares hit previously. However, this comment may merely be evidence that 'random' and 'randomness' are terms to be avoided [although current usage likely means we must (reluctantly) retain 'random variable'].

11. Uncertainty in Statistics

Variation is ubiquitous in the real world. More specifically, variation among response and explanatory variate values of the elements of a population is a source of *uncertainty* in answers to (statistical) questions based on a *sample* selected from the population. This is an example of the more general difficulty of *inductive* (as opposed to *de*ductive) inference – reasoning from the particular (like a sample) to the general (like a population). Thus, in the presence of variation, uncertainty is *unavoidable* in an answer based on *incomplete* information (see Statistical Highlight #87) – *we* characterize this uncertainty as *limitations* on answers and our six error categories (as described in Statistical Highlight #6, for instance) focus attention on statistical methods which try to make the severity of the limitations acceptable in the Question context.

A distinction between (ubiquitous) variation and uncertainty in an answer is that the latter implies the presence of a *sentient* being to frame the Question; however, it is easy to confuse variation and uncertainty. For example, explanations, suitable for a non-technical audience, of Heisenberg's Uncertainty Principle may talk more about variation than about uncertainty, to the point where a reader might wonder if a better name would be Heisenberg's Variation Principle.

In statistics, we quantify uncertainty in terms of behaviour under repetition, which involves the regularity that is a theme of earlier discussion in this Highlight #94. However, we must not allow this to obscure short-term unpredictability which is the source of limitations on the answer(s) from a *particular* investigation. Once the data have been collected in the Execution stage of the FDEAC cycle, the amount of information they contain, relevant to answering the Question(s), is *fixed*. The following Analysis stage can then explоint this information more or less effectively, depending on the statistical method(s) used, to obtain and convey the Answer(s). Obviously, the ideal is adequate data (obtained with minimum expenditure of resources) and correct and clear analysis. However, an *in*adequate analysis of adequate data is typically (much) less resource intensive to redress than data that are *in*adequate to properly address the Question(s), when the Plan must be redeveloped [or even the Question(s) reformulated], followed by resource-intensive repeating the investigation. This is the danger of undue emphasis in statistics teaching on mathematics and statistical methods of *analysis* [discussed overleaf (∗) near the top of page HL94.11] at the expense of stressing the essential roles of Question formulation and Plan development in the first and second stages of the FDEAC cycle.