THE CONCEPT OF DISTRIBUTION

CHRIS WILD
The University of Auckland, New Zealand
c.wild@auckland.ac.nz

ABSTRACT

This paper is a personal exploration of where the ideas of "distribution" that we are trying to develop in students come from and are leading to, how they fit together, and where they are important and why. We need to have such considerations in the back of our minds when designing learning experiences. The notion of "distribution" as a lens through which statisticians look at the variation in data is developed. I explore the sources of variation in data, empirical versus theoretical distributions, the nature of statistical models, sampling distributions, the conditional nature of distributions used for modelling, and the underpinnings of inference.

Keywords: *Frequency distributions, Statistical models; Sampling distributions; Statistical inference; Types of distribution; Variation*

1. INTRODUCTION

There are aspects of statistics that are so basic to the way we think in the subject that no one abstracts, enunciates and examines them. We encountered this phenomenon frequently in conducting the research for Wild and Pfannkuch (1999). It is not a problem for the statistical practice of professionals since they have long since been successfully encultured into these ways of thinking. It may well, however, be a root cause of some of the problems we face in statistics education. "Variation" was one of these unenunciated givens until quite recently and still is for many communities of statisticians. "Distribution" is another fundamental given of statistical reasoning. I can find a great deal written about specialized usages and definitions of "distribution" but almost nothing about "distribution" itself as an underlying conceptual structure. For example, Wiley's massive 16 volume *Encyclopedia of Statistical Sciences* does not contain an entry for "distribution" as an entity although it contains over 300 different sections in which "distribution" appears in the title.

The main aim of this paper is to try to explore for teachers and statistics education researchers where the ideas of "distribution" that we are trying to develop in students are leading to, and where they are important and why. We need to have such considerations in the back of our minds when designing learning experiences. They are a logical precursor for a planned educational development; a platform upon which the educational "when?," "in what order?," "by what means?" and so on, can be built. Our journey towards an understanding of "distribution," and the need for concepts of distribution, begins with the pervasive nature of variation.

Section 3 of Wild and Pfannkuch (1999) was entitled "Variation, randomness and statistical models." The genesis of that story was, "In the beginning was variation." Variation is an observed reality detectable in all systems and entities. It is, in a word, omnipresent. A statistical response is generated when the variation we have to deal with

Statistics Education Research Journal, 5(2), 10-26, http://www.stat.auckland.ac.nz/serj

in pursuing a real-world goal is not completely predictable at levels of precision that are of practical importance and we have given up, at least temporarily, on the ability to understand differences between individuals at a level that might make them predictable. The statistical response is to investigate, disentangle and model patterns of variation in order to learn from them. We will see that the notion of "distribution" is, at its most basic, intuitive level, "the pattern of variation in a variable," or set of variables in the multivariate case. Thus the notion of "distribution" underlies virtually all statistical ways of reasoning about variation. So it is particularly fitting that the first special section of the *Statistics Education Research Journal* (Garfield and Ben-Zvi, 2005) had the theme of "variation" and this, the second special issue, has the theme of "distribution."

Statisticians look at variation through a lens which is "distribution" (Figure 1). Provided "variation" is in the background of our thinking, we are looking through the "distribution" lens as soon as we look at our data in any way that sets aside case labels. Setting aside case labels is no small matter, however. There has been a good deal of work about how children at elementary and middle school levels relate to data. Bakker and Gravemeijer (2004, p. 147) write that such students "tend to conceive a dataset as a collection of individual values instead of an aggregate that has certain properties." What is important and interesting to children is the particular. Case labels (e.g., the names of people) inform us that a particular data record describes a specific entity, often a person. It is a very big step indeed from this to thinking about data in aggregate terms.

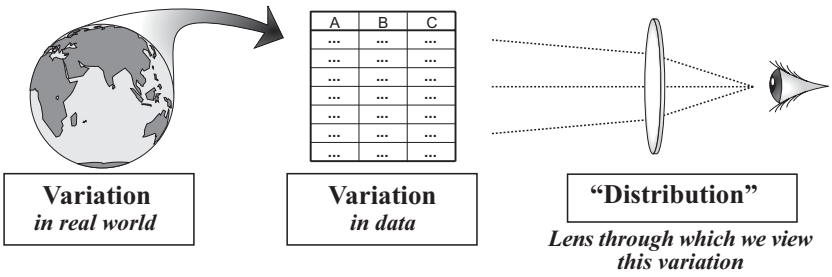


Figure 1. "Distribution" as a lens

In statistics we are seldom interested in a dataset as a collection of separate snapshots of particular individuals taken in a particular way at a particular instance in time. Rather we look at data to learn more widely applicable lessons. These lessons are not, we believe, to be found in the individual data points themselves but in patterns discernible in the dataset as a whole. So we put aside (temporarily ignore) the links between data points and individuals as distracting detail in order to better focus on patterns. When case labels are set aside individuals with identical values for the variables of interest become indistinguishable so that, without any loss of information, we can reduce the data to a set of distinct values and their corresponding frequencies, that is, to a frequency distribution. All of the information about patterns of variation is in the (typically multivariate) frequency distributions. All summary statistics and almost all the graphs we look at are summaries and graphs of frequency distributions. We use them to discover and describe aspects of the patterns in the variation contained in the frequency distributions. We convert frequency distributions into relative-frequency distributions to facilitate the comparison of batches of data (e.g., to compare data from different subgroups) containing different numbers of observations.

Where does the variation we see in data come from? There is typically real variation in the systems we are investigating and this is inevitably overlaid with additional variation induced by the observational process as in Figure 2. Why do we summarise and model patterns of variation? Primarily we do it for the purposes of prediction, explanation or control; that is, in order to be able to make better predictions, better understand the mechanisms generating the data, or to enable us to change the pattern of variation in the system in the future, at least to some partial but useful extent such as by reducing mortality rates.

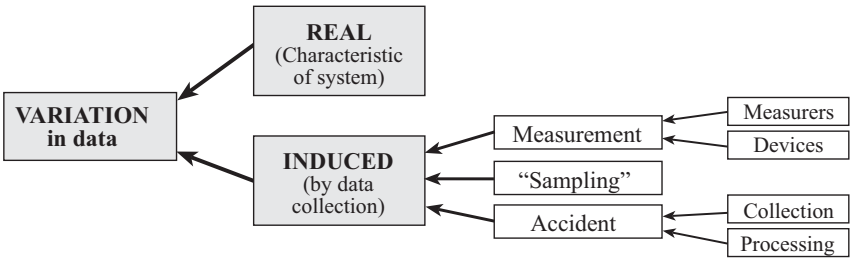


Figure 2. Sources of variation in data

Section 3 of Wild and Pfannkuch (1999) went on to discuss how statisticians look for sources of variability by looking for patterns and relationships between variables and in particular for those patterns that are likely to persist. It talked about explained and unexplained or residual variation. The majority of the section discussed "the quest for causes" and I don't want to touch on that here (except to promulgate "variation causes statistics"!). It concluded with the following: (1) variation is an observable reality; (2) some variation can be explained; (3) other variation cannot be explained on current knowledge; (4) random variation is the way in which statisticians model unexplained variation; (5) this unexplained variation may in part or in whole be produced by the process of observation through random sampling; (6) randomness is a convenient human construct which is used to deal with variation in which patterns cannot be detected.

We look for regularities or patterns in the observed variation and those that we believe, considering what we see in the data and what we understand about the mechanisms generating the data, are likely to be real and not ephemeral correspond to "explained variation." Unexplained variation, or "noise," is what is left over once we have "removed" all such patterns. It is thus, by definition, variation in which we can find no patterns. We model unexplained variation as being generated by a random process, implicitly if not explicitly. The simplest such models are regression models. We are papering over, at this point, a rather large crevasse which is the difficulty in deciding whether an apparent pattern in our data is likely to be a persistent characteristic of the process generating the data, and thus form a structural element in our model, or ephemeral and should be swept up in random elements of a model.

There is an old saying that goes, "If it looks like a duck, walks like a duck and quacks like a duck, then it is a duck." If it looks/walks/quacks like a duck, the statistician will use the inferential reasoning appropriate for ducks, despite having no real assurance that this bird actually has duck DNA. When modelling unexplained variation, because it looks random when viewed in any of the ways we have devised for inspecting it, we will draw the inferences that we know would be appropriate if it was in fact randomly generated. We do this because we do not know any better ways of proceeding (and don't believe anyone else does either). For further discussion, see Section 3.4 of Wild and Pfannnkuch (1999).

Having established "distribution" as a lens through which we view variation in data and explored the nature of explained and unexplained variation we will now start looking at distinctions between types of distributions that draw on these ideas.

2. EMPIRICAL VERSUS THEORETICAL DISTRIBTIONS

2.1. INTRODUCTION

In an effort to understand better how statisticians use "distribution," I pointed Google at a number of sites including the American Statistical Association (ASA) where it searched the pages of the *Journal of the American Statistical Association*, the *Journal of Statistics Education*, other ASA journals, and many other resources. The adjectives and other qualifiers that I found used with "distribution" are collected in Appendix 1. By far the most common usages fell into two classes, "named theoretical distributions" (e.g., normal, binomial, …) and "the distribution of …" referring to the empirical or frequency distribution of some particular measured quantity, so that will be our starting point.

The distinction that underlies discussions of *empirical* versus *theoretical* distributions is between the variation we see in our data and a potential model for the process that gives rise to that variation (Figure 3). The empirical, *frequency* or observed distribution of our variable(s) contains the variation that we can see directly in our data. There is no inferential component, just a description of what exists in the data. When we move on to try to learn wider lessons from features seen in the current dataset, we conceive of unexplained variation present as having been generated by some unknown distribution. We often refer to this as the "true" or "underlying" distribution even though it is almost always a conceptual entity. When we use a full parametric model in our analysis we choose some named parametric distribution, such as the normal distribution, which we then assume to be what generates the data. This is the *theoretical* distribution, which describes or defines a probability model.

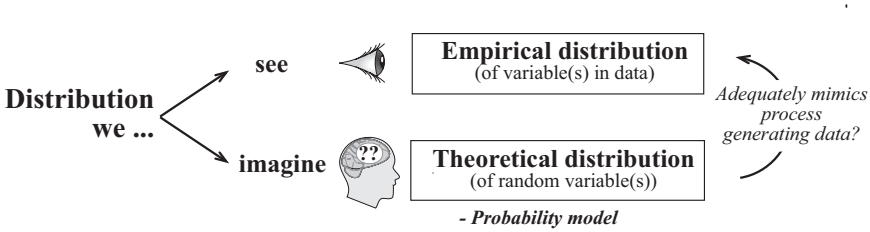


Figure 3. Empirical versus theoretical distributions

We have hundreds of humanly invented distributions for such purposes. In certain application areas, experience has shown that certain distributions are useful, but there is no way of ever knowing that our data are being generated from some particular distribution. So we never really believe our assumed theoretical distributions. The best we can hope for is that the act of sampling from the assumed theoretical distribution adequately mimics the most important features of the process which generated our data. Our lack of trust in the theoretical distribution leads to considerations of "robustness" and "goodness of fit." That is, we would like to use inferential procedures that are comparatively insensitive to departures from distributional assumptions (robust) and we want to avoid using a theoretical distribution for inference that demonstrably does not "fit" the data – by which we mean that the distribution would be unlikely, in some sense, to produce the dataset we have in hand. To have any hope of making sense of this

modelling process, students need to experience the behaviour of data which are generated from truly random sources (or as close as we can get to that) and lay that alongside real data. This is the crux of "connecting chance and data." I will expand on this point in the next subsection.

Where do *outliers* fit? Outliers are observations we suspect are not being generated by the process which is generating the bulk of the data, but by a different (e.g., gross error) process. When we detect outliers we go back to the case labels hoping that there is some additional information we can uncover about that case that might help us understand why it appears so different. For example, we may be able somehow to determine whether the outlier is an error that can be corrected or removed.

2.2. "HEIGHTS ARE NORMALLY DISTRIBUTED"

At risk of belabouring points, I will now approach the ideas in Section 2.1 from another direction. How can we understand a statement like "heights are normally distributed"? We should not understand it in an absolute, literal sense because the statement is far more precise than anything we could ever actually know. Usually we are using "heights are normally distributed" in a loose descriptive way. The shape of the empirical distribution of the heights that we have seen (in whatever context) looks as though it is reasonably well approximated by the probability density curve of some particular normal distribution. We may make a leap of faith by believing that the approximation would still be good if we could look at the empirical distribution of heights from everyone in the parent population from which the heights we have seen have been drawn. We could make an even greater leap by thinking that this is probably also the way it would turn out if we looked at heights of people drawn from some other population that we have not yet investigated.

If we add the idea or reality of sampling at random from a population where the height distribution is well approximated by a normal distribution, then it follows that the behaviour of the data we get from sampling people and measuring their heights should be almost indistinguishable from the type of data we would get from taking random draws from a normal distribution. That latter behaviour can be investigated directly mathematically or via simulation.

If we make the assumption that our data on heights have been sampled from a Normal probability model then inferential statements (e.g., a confidence interval for the mean of the heights population distribution) follow from statistical theory as a consequence of that assumption. This is analogous to mathematics where, if one takes a set of conditions as holding true (axioms), then many other statements deduced as a logical consequence of these initial axioms (the theorems), must also hold true.

Some of the distributional leaps of faith in the first paragraph may be informed by a nonsignificant test of normality for our height data. But how much does this tell us? It tells us only that we cannot rule out the possibility that sampling variation alone may have produced the degree of "departure from normality" that we see with these data. Experience shows that, in virtually every situation, any theoretical distributional assumption we care to make will be shown to be implausible given enough data. What we are doing is never about the assumed theoretical distribution being right. It is only ever about the assumed theoretical distribution being a close enough approximation so that the methods of drawing inferences that follow from the assumptions we make are not misleading in any important way. This brings us back to robustness and goodness of fit as discussed in Section 2.1. *We make distributional assumptions in order to come up with*

methods of drawing inferences from data that still work when those distributional assumptions are not quite right.

There is a very understandable desire to drive the teaching of probability models and distributions using real data because dice, coins, and so on are boring, even irrelevant. This runs into the problem discussed above. We can never know that any specific set of real data has been generated from any specific probability model. At best we can believe that the model is a good approximation. Probability models are abstract constructs that are used to model real-world behaviour. Their successful operation stands on two legs. The first leg consists of understanding the abstract construct that is the model, the sort of "data" the model generates, and how we reason inferentially in that idealised environment. The second leg consists of seeing the parallels that suggest to us that the model may provide a reasonable approximation to a given reality, of applying the model-based reasoning, and then of interpreting the results in terms of the original context. For most of the mental connections that have to be built in order to understand the model and the nature of its random behaviour, real-world context is simply a distracting irrelevance. That is not the place of real-world context. Interaction with context occurs in the recognition of model applicability, the interpretation of model parameters and the interpretation of any inferential statements that follow from applying the model-based reasoning.

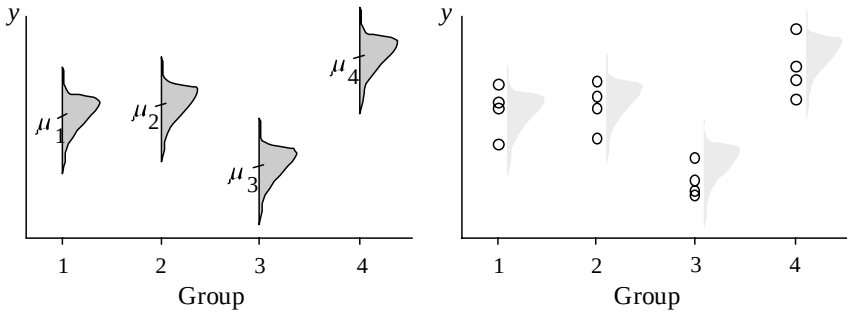
As a non-traditional illustration, what students are experiencing in the fascinating basketball environment described by Prodroumou and Pratt (2006) is the stochastic behaviour of simulated "data" generated by a statistical model. While students do not directly learn anything new about basketball, by adjusting model parameters they can make the behaviour exhibited by the simulated environment (i.e., the statistical model) feel a lot like that of basketball. They can play with strategies that affect their performance in the simulated game and if they believe that the simulation gives an approximation that is close enough in key features to basketball they should then be willing to transfer some of the lessons learned in the simulation environment to the actual game.

2.3. MORE ABOUT DISTRIBUTIONS AND MODELS

We now explore some complications neglected in the discussion in Section 2.1. When we choose "a" theoretical distribution as a model for some variable, typically we are actually referring to an assumption that the true distribution is an unknown member of a parametric family of distributions such as the Normal(μ,σ^2) family. Here assigning different values to the parameters μ and σ^2 gives rise to different distributions within the family and we make statistical inferences about the unknown "true" values of the parameters that "produced the data."

Beyond the simplest models, we do not just specify "a distribution." We actually build a construct using structural and random elements where each random element has a distribution. The simplest models of this form are the one-way analysis of variance model depicted in Figure 4, which underlies traditional inferential methods for comparing groups, and the simple linear regression model depicted in Figure 5. In Figure 4 the shapes are little normal curves coming up out of the page. The normal distribution for the *y*-values in group *i* is centred at μ_i. Under this model, "observations" belonging to the *i*th group are generated by sampling from a normal distribution with mean μ_i and some variance σ^2, which is the same for all of the groups. This is represented on the right hand plot, which also retains a "ghost" of the generating distribution. The model generates a type of pattern that we often observe in real data when we are trying to compare groups

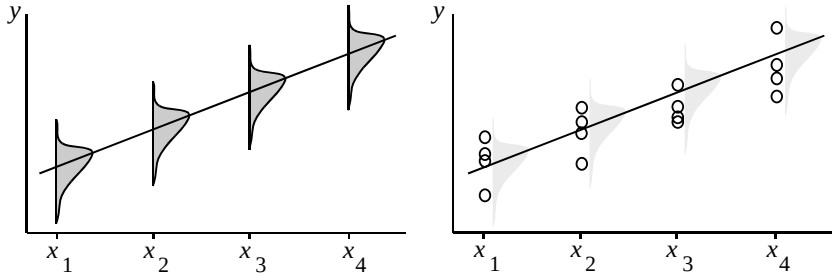
and thus forms a model for the mechanism generating such data that we can often apply in practice. In reality, the values of μ_i and σ^2 are unknown. Standard statistical inferences include testing for, or finding confidence intervals for, differences between the true group means.



(a) One-way Anova model (b) Data sampled from the model

Figure 4. The one-way analysis of variance model

The simple linear regression model in Figure 5 is essentially the same except that the true means μ_x when plotted against the x-value at which an observation is taken are constrained to lie on a line. The structural part of this model is the linear relationship between x and the mean value of y. The random part is the distribution of y-values taken at a given x around that the mean and that is what generates the observed scatter about the linear pattern.



(a) The simple linear model (b) Data sampled from the model

Figure 5. The simple linear model

Variation seen in even very simple data structures stems from a variety of sources (e.g., person-to-person, measurement, or occasion-to-occasion). There is a need to be able to think in quite sophisticated distributional ways to tease these things out. Hierarchies of random components (multilevel modelling) can be very helpful here. Luckily, in many commonly encountered problems it is not necessary to do so. Naïve approaches that sweep the subtleties under the carpet are actually valid. Suppose, for example, we want to compare the blood pressures of a drug-treated group and a control group on placebo. People do not have "a blood pressure." At the very least there is person-to-person variation in the levels of their average blood pressure, occasion-to-occasion variation in actual blood pressure of the same individual, and measurement error is a third source of variation adding to the other two. The variability of blood-pressure readings seen within each of the two treatment groups is the result of all of these sources. Nonetheless, with only a single observation per individual (and admittedly under certain idealised

assumptions) a 2-sample *t*-test, or confidence interval, for a difference in mean levels is still a valid analysis regardless of whether we have all of these sources of variation operating or if only person-to-person variation was operating. What differs is how we interpret the within-group variances. We would not raise the complications of multiple sources of variation for inferential beginners to avoid cognitive overload but suspicions about them might well cause unease in some.

Many complex models for processes involving time and space are built up in terms of chains or hierarchies of conditional distributions. For building models for processes evolving in time, for example, we often build up probability models conditionally by thinking in terms of what might happen next given the history of the process up until that point. This is a way of conceptualizing that permits prediction and also enables us to cope with data features like censoring.

With Bayesian inference we push the conceptual envelope out still further with the idea of describing the state of knowledge (prior to collecting the data) about parameters in a distributional model in terms of distributions called *prior distributions*. A Bayesian treatment of one-way analysis of variance, for example, would include prior distributions for all of the μ_i's and the variance σ^2. Inference proceeds by updating these prior distributions using information in the data to form corresponding *posterior distributions* intended to encapsulate the new data-informed state of knowledge.

2.4. ALL DISTRIBUTIONS ARE CONDITIONAL

All distributions we work with are really "conditional distributions." This is not to say that we need complicated conditional probability ideas to think about them, just that they apply to particular subpopulations or systems operating under particular conditions or "settings" or to a particular time. We want to plant the idea that as conditions (or the groups we look at) change, the pattern of variation in an outcome variable often changes too and that we can learn useful things when we can quantify or otherwise describe the nature of those changes. If we can do this there is useful predictive information in such things as group membership and an impetus is given to trying to understand why the patterns might change. The regression problem can be conceived of as an investigation into how the distribution (pattern of variation) of a response variable y changes as the setting (x) changes. Group comparisons (two-sample, analysis of variance, etc.) can be conceived as an investigation into how the distribution (pattern of variation) of a response variable changes as we move from subpopulation to subpopulation (group to group) as shown, in an idealised way, in Figures 4 and 5. In the models depicted in Figures 4 and 5, all that changes about the distribution of y–values as group membership changes (Figure 4) or x changes (Figure 5) is confined to the mean level of the response. Spread, shape and everything else remains identical. Of course, even if this was true of the mechanism generating the data, in any observed dataset all of the features of the empirical distributions will still differ from group-to-group at least to some extent.

Regression and analysis of variance problems are not usually presented at this level of generality. The emphasis in most textbooks is not on how the distribution changes but on how the mean changes. Why this emphasis just on means? There are many reasons. One is a desire to look at the simplest feature of the distribution first. Then there is the historical influence of having well-worked out theory for simple models in which the mean is the only thing that changes as x changes (or as we move from group to group). Additionally, the parsimony principle (or Keep-It-Simple-Stupid principle) leads us to model only changes in mean unless the data forces us to do something more complicated. Other characteristics are much harder to make inferences about. For example, normal

theory-based inferences for means are quite robust but those for spreads are extremely sensitive to departures from normality assumptions. As an indicative convention, the more "detailed" the feature being compared, the more data we require to usefully characterise or compare it.

2.5. SAMPLE DISTRIBUTION VERSUS POPULATION DISTRIBUTION

For beginning students we usually introduce the distinction between empirical and theoretical distributions gently via the distinction between the *sample distribution* and the *population distribution*. More precisely this is the distinction between the distribution of values for a variable for individuals represented in our dataset versus what that distribution would be if we had data on everyone in the population. Beginning this way is consistent with our desire in teaching to move sufficiently slowly from the concrete to the conceptual so that students do not drown in subtlety. Distributional models for data from processes are necessarily conceptual and immediately raise all sorts of difficult questions, for example, about the stability of the process through time and space and about dependencies. With data from a population, however, we can think in much simpler terms, namely of sampling from a large set of individuals at one point in time, and measuring one or more characteristics on each individual selected.

In practice, however, nothing is ever quite that simple, quite that "concrete." The really concrete, (real finite population measured once with one device at one time by one person in one way) is not really of interest to anyone because the quantities of interest are confounded, at the very least, by measurement-process variation. What we see is not exactly what is there. As soon as we allow for a contribution of the measurement process to the variation present in the data we are immediately transported from a manageable easily understood world to a world where data are generated by sampling from a conceptual population (an imagined construct) or are generated by some sort of random process (see Konold & Pollatsek, 2004). Urban myth has it that mediaeval mapmakers alluded to dangers lurking beyond the borders of the known world with the phrase "Here There be Dragons." Our maps of the statistical/inferential world made for beginning students need to be inscribed very carefully for teachers with "Here There be Dragons" underlined with "These Dragons be Real."

3. SAMPLING DISTRIBUTIONS

Next in importance, after the empirical and theoretical distributions of observations, are *sampling distributions* (see Figure 6). The former two relate to the *unit-to-unit variation* that we can see within a study or dataset and to a model for the generation of that unit-to-unit variation. (I prefer to personalize this and speak in terms of individual-to-individual variation.) Sampling distributions relate to *study-to-study variation* in estimates or statistics (e.g., sample means, proportions, regression slope estimates and *t*-statistics) which cannot be demonstrated from any particular study because each research study provides only one study-level data point. It is most accessibly introduced to students, I believe, in terms of the sampling variation in a parameter estimate, for example, of a population mean or proportion. Statistics educators now have a very good array of complementary ways of enabling students to experience the sampling variation generated by the process of "conduct a study and calculate an estimate" (see Chance, delMas & Garfield, 2004, pp. 294-297). This sampling variation can be modelled using either a (theoretical) probability distribution deduced from the distribution used to model the unit-level data or an asymptotic (large-sample) approximation, or it may be simulated

using a resampling technique be it bootstrap, jackknife or permutation depending upon the situation and the analyst's taste.

The main priority with sampling distributions is to get across the idea that estimates and other statistics change every time we do a new study even if we perform each study according to exactly the same protocols. Properly appreciated, this becomes the prime motivator for the need for inferential methods which incorporate uncertainty, be they significance tests and confidence intervals or Bayesian. A second priority is the Central Limit Theorem for means which lays the groundwork for commonly used inferential techniques for a range of simple, but common, situations.

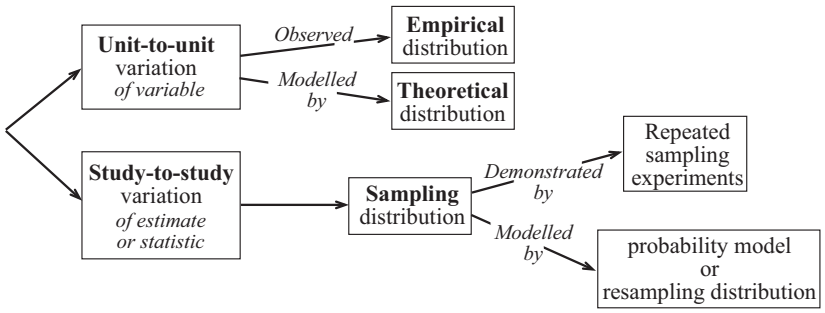


Figure 6. Incorporating the sampling distribution

Although the expression "sampling variation" appears often in the statistics education literature it appears very rarely in the statistics literature. There is no article on sampling distributions in the *Encyclopedia of Statistical Sciences* and of the over 300 section headings that contain the word "distribution" the phrase "sampling distribution" appears in only three. It is a background concept that underpins much of what we do in statistical inference but once the idea has been established it is seldom explicitly referred to. The adjective "sampling" is either dropped, to be inferred from the context, or it may appear in other guises such as the *null distribution* (of a test statistic), which is the sampling distribution that the test statistic would have if the null hypothesis was true.

4. COMPARISONS CURE UNIVARIATITIS

Teaching about the features of distributions for beginners tends to be in the context of a single variable, that is, in a univariate setting. All too often this has led to students being fed, year after year, a constant diet of univariate data and contrived univariate situations. I plead with teachers to move on to multivariate notions such as comparisons between groups and relationships between variables as soon as the most basic foundations have been laid. This is necessary to avoid infecting students with the dread disease univariatitis which is notorious for causing its victims to experience sensations of drowning in irrelevance and, ultimately, death by boredom. We may have to keep revisiting the univariate world but should take extreme care not to end up living there.

One important reason for using multivariate data early is that it gives a time-efficient environment in which students can themselves generate interesting questions to investigate using data, for example, by making interesting comparisons or investigating possible relationships. As pointed out in Wild (1994, p. 164), "not only is question generation arguably the most important part of the investigative process, the bubbling up of questions from an awakened curiosity provides much of the excitement of investigation."

Single distributions and the features of single distributions are seldom of interest in and of themselves. Interest generally lies in changes in these features, between places or groups, or over time. So why do we spend so much time working with single univariate distributions? Our main purpose is to lay the conceptual ground work that facilitates thinking about comparisons and relationships but traditional statistics teaching spends far too much time on it. We start with data and talk about centre, spread, modes, gaps, clusters, skewness, quantiles (particularly medians and quartiles), outliers, and other words are also starting to be used out of concerns about language being child-friendly and descriptive (e.g., "spreadoutness," "clumps," and "bumps"). This all too easily turns into what my colleague Matt Regan pejoratively terms "name calling." Let us put the simple data features to work in comparisons before we start naming and worrying about more detailed data features. Useful inferences about the latter are much less common in practice and much less reliable as well. So as soon as we introduce ideas like centre and spread we should put them straight to work in making some real and interesting comparisons – having visited the dull, grey, univariate world we need to bring the learning straight back into the vibrant real world. The same applies for notions like skewness. The fact that data for some variables are severely skewed is interesting mainly because data on other variables are not (another type of comparison) and because of practical implications of distributional shapes.

One of the many things we want students to be able to do when looking at plots of their data is to react to and wonder about causes for "the unexpected," particularly outliers – things that fall beyond "the expected pattern of variation." In order to do this students need some ideas about what to expect. A good place to start is the patterns of variation produced by sampling from a normal distribution or a finite population in which the characteristic of interest is approximately normally distributed. Particularly with small to moderate samples, the extent of what we might think of as "non-normal behaviour" present in data generated from a normal distribution can be astounding. Meaning should only be sought in those features of the data that correspond to features of the parent population or other mechanisms generating the data. Because exploratory data analysis is seldom coupled with exploration of models and random behaviour, many of the features beginning students point to, name and ponder causes for (gaps, clumps, outliers, skewness, bimodal behaviour, etc) are within the threshold of random error.

A recent innovation for the beginnings of inference introduced explicitly by Bakker and Gravemeijer (2004, pp. 158-165), but also used by others, for example, Konold and Pollatsek (2004, pp. 172, 180, 193), is the mind game for children of "growing the sample" which is basically concerned with conjecturing about what we might expect to happen to a display if "we added more people." In our terms, a data feature is meaningful only if it would still be present if we grew the sample substantially. For example, a gap would not be filled in, or apparent clusters would not coalesce. We move beyond name calling to statistical thinking when we can relate the features that we can see and name in our dataset, and believe will persist, to what we know about the world in order to arrive at some level of real-world insight, however small. We may, as a simple example, identify two clusters in a distribution and through further detective work determine that they are composed of identifiably different classes of individuals.

With categorical data, the most important reasons for working with relative frequencies (equivalently proportions or percentages), such as in relative frequency tables and resulting bar graphs, is to facilitate the comparison of datasets of different sizes, and to form a bridge to probability. With continuous measurement data, the real reason for teaching standardized histograms in which proportions are represented by areas is to lead in to the idea of probability density and density curves. This form of standardisation also

permits comparison of datasets which have been summarised using different class intervals and to display a single set of data that has, for some other reason, been summarised using class intervals of different width. In practice, however, the need to do either of these things is so rare that I would never give it class time.

Some in the statistics education research community have found proportions of a sample below/above cut points (e.g., proportions of girls and boys with a height above 120 cm) provide a child-friendly introduction to the making of comparisons between groups when the response variable is continuous. It appears that this is something that many children do almost spontaneously. The reason we seldom see it in more advanced treatments is because the choice of cut point tends to be arbitrary and because this method of making comparisons is statistically inefficient. For example, more data is required to demonstrate a significant difference between groups this way than by comparing means. Statistical inefficiency does not provide a convincing argument against beginning students engaging with data in a way that is natural to them, however. It is much more important that they are enculturated to engage. Moreover, there are important areas in which the cut-point method is used, at least for communication purposes. Medical reporting often employs 5-year survival rates, for example.

5. DISCUSSION

The ultimate goal of statistical investigation is learning about some external reality and this involves forming and updating models of this context reality. In applied statistics there are three main elements that are brought together: current understandings of the context reality, data, and the use of statistical models and knowledge to guide how we collect data and learn from our data (understandings). Figure 7 attempts to represent the interrelationships.

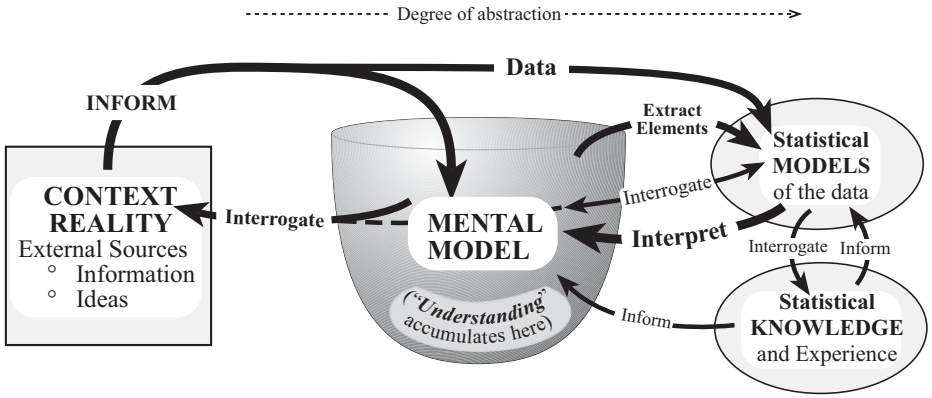


Figure 7. Learning via statistics

The need for statistics flows from variation, particularly the presence of "unexplained variation," in data. The statistical response to variation is to investigate, disentangle and model patterns of variation in order to learn from them. Virtually all of the ways statisticians do this involve looking at data through a lens which is distribution. While labels are interesting to children they have to learn to set aside labels and move beyond "who is this?" to start seeing and focussing on the patterns of variation and then to thinking about what aspects of these patterns might be expected to persist more generally. As Rubin et al. (2005) state, "aggregate views are preferable, as they are required to look beyond the data towards making inferences about the underlying populations or processes

represented by data samples." When case labels are removed data records with identical values for the variables become entirely exchangeable and we are left with frequency distributions. The graphs and summaries we use are ways of looking at, summarising and conveying aspects of the information present in these distributions. Fundamentally, the notion of "distribution" is the pattern of variation in a variable (or set of variables).

The operation of the thinking processes represented in Figure 7 rests heavily on the interplay between the behaviour of our data and understanding the stochastic behaviour of potentially useful statistical models. To do this well requires bringing together the two elements of experience with exploratory data analysis and experience with the stochastic behaviour generated by models. Empirical (frequency) distributions tell us about data behaviour, whereas theoretical distributions are critical conceptual building blocks for statistical models. Put another way, the distinction underlying empirical versus theoretical distributions is between the variation we see in our data and a model for the process that generates that variation. We conceive of unexplained variation present as having been generated by some "true" or "underlying" distribution. In a full parametric analysis we assume these distributions are unknown members of a known, named parametric family of distributions. The idea of "population distributions" may paper over some complications for beginners but the paper is usually very thin.

Whereas empirical and theoretical distributions of observations relate to within study (or within dataset) variation that we can imperfectly see, sampling distributions relate to study-to-study variation in estimates or statistics which cannot be seen from any particular study because each study provides only one study-level data point. Sampling distributions motivate the need for and are a component of the development of statistical inference.

All distributions are conditional in the sense that they apply to particular subpopulations or systems operating under particular conditions or "settings" or to a particular time. The regression problem can be conceived as an investigation of how the distribution of a response variable changes as the setting (x) changes and group comparisons can be conceived as an investigation of how the distribution of a response variable changes as we move from group to group.

Because distributions are such a fundamental component of statistical reasoning our main goal should not be, "How do we reason about distributions?" but "How do we reason with distributions?," moving from a world where individual atoms are what is interesting to reasoning using aggregates. As Watson (2005) writes, children are beginning to learn about distributions from an early age starting when they first create pictograms of favourite fruits or modes of transport. They are not told, and do not need to be told, that they are learning about "distribution." Students typically first encounter summary features of distributions such as means, medians and even interquartile ranges long before they have any but the vaguest idea of "distribution." We look at graphs of distributions long before we develop the notion of distribution. Indeed our more complex notions of distribution and the nature of various features of distributions draw heavily upon the behaviours that have already been seen exhibited in graphs of data.

So do students need to be able to form and articulate a concept of distribution to be able to operate in a statistical way? Or, to steal from Nike, can students "Just do it" using graphs, summaries and an intuitive appreciation of variation? My feeling is that an explicit notion of distribution is not needed until we want to motivate, understand and then use probability models. Although distribution is the second foundation stone on which statistics is built ("variation" is the first), what is critical for early learners is much less, *"What is 'distribution'?"* than, *"How are my data distributed?"* and beginning to answer that question using appropriate graphs and summaries. One of the usual English-

language meanings of the word "distributed," taken from the *Oxford English Dictionary*, "is spread or disperse(d) abroad through a whole space or over a whole surface." The first step is that measured characteristics of individuals (e.g., heights) are not identical – their values are distributed across a range – and that we can learn useful things by looking at how they are distributed. "How are my data distributed?" points someone like me in the right direction but does it speak to the variety of students that might experience it? We can be very grateful we have committed researchers in statistics education who are prepared to pick up questions like the above, and the issues raised below, and move the discussion beyond conjecture and anecdote.

It is my belief that we should be forming mental habits in which raw data (numbers) prompt students immediately to reach for pictures of that data, or as Moore (1991, p. 426) says, "a structure of thought that whispers, 'Variation matters … Why not draw a graph?'" Summaries should be related conceptually to these pictures. Pictures of data, of distributions, do not need to be conventional pictures though they should converge to them over time as statistical knowledge develops; there are good reasons why conventional pictures have taken hold. Someone's student somewhere at some time may very well come up with something startling and new which should inform the way everyone else does graphics but we must expect this to be rare. A trap for teachers is the presumption that conventional pictures are easy to read. Tools which are so transparent to the initiated (nothing to teach, it's completely obvious, how could anyone *not* get it?!) can be quite opaque to beginners. Students need also to learn that there is not just one correct picture, that we can form a better overall view of reality by using an array of different pictures that better highlight different features of the data. Section 4 of Gould (2004) provides good examples of this and also how the insights so obtained feed into the development of statistical models for the data.

The key drivers for successful statistical practice, and thus the most critical elements to be instilled by statistics education are three propensities: the propensity to collect data that usefully addresses the question of interest, the propensity to question the applicability of data to the problem in hand, and the propensity to seek meaning in data. Everything else is about how to act on these propensities.

Most of the papers at STRL-4 and in this special issue deal with students' engagement with empirical distributions, their features, and with comparisons of these features between groups. When features like location shifts between groups show up in a set of boxplots, for example, the following questions are never far off. "But does it actually mean anything?" "If we did it again would it come out much the same? Or could the order of the groups even be reversed?" Instantly we are transported to the realm of inference.

The inferences beginning students are able to make are necessarily informal, but therein lies the rub. There are great difficulties with informal inferences as Pfannkuch (2006) discusses. Assessment of "significance" balances the three factors: effect size, variability and sample size, in a very complicated way. Sets of standard boxplots that look identical, except for being based on different sample sizes, must be interpreted differently (notched box plots, Garret & Nash, 2001, provide a workaround). It may well be that there are no easy answers. There were some very sound imperatives that drove the development of our formal schools of statistical inference! As statistics educators we need to encourage our students into the mental habit of continually seeking meaning in data, which includes trying to make inferences, even using inadequate tools. The focus for the next SRTL Research Forum (SRTL-5 in 2007) and, I hope, a future special issue of *SERJ*, is informal ideas of inference. I look forward to the results with extreme interest.

ACKNOWLEDGEMENTS

The author is very grateful to Maxine Pfannkuch, Chris Reading and the participants at SRTL-4 for helpful suggestions and discussions.

REFERENCES

Bakker, A. and Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chance, B., delMas, R. and Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J., and Ben-Zvi, D. (2005). Reasoning about variation [Special Section]. *Statistics Education Research Journal*, *4*(1), 27-99.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ4(1).pdf]

Garrett, L. and Nash, J. C. (2001). Issues in teaching the comparison of variability to non-statistics students. *Journal of Statistics Education*, *9*(2).
[Online: www.amstat.org/publications/jse/v9n2/garrett.html]

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, *3*(2), 7-16.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Gould.pdf]

Konold, C. and Pollatsek, A. (2004). Conceptualizing an average as a stable feature of a noisy process. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 169-199). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Moore, D. (1991). Statistics for all: Why? What and how? In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics (ICOTS-3)*, Dunedin, August 1990, Vol. 1 (pp. 423-428). Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M. (2006). Comparing boxplot distributions: A teacher's reasoning**.** *Statistics Education Research Journal*, *5*(2), 27-45.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Pfannkuch.pdf]

Prodroumou, T. and Pratt, D. (2006). The role of causality in the Co-ordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69-88.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Prod_Pratt.pdf]

Rubin, A., Hammerman, J. K., Puttick, G., and Campbell, C. (2005). Developing models of distributions using tinkerplots. In K. Makar (Ed.), *Reasoning about distribution: A collection of current research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy,* Auckland, 2-7 July 2005, [CD-ROM, with video segments]. Brisbane, Australia: University of Queensland.

Watson, J. (2005). Developing an awareness of distribution. In K. Makar (Ed.), *Reasoning about distribution: A collection of current research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy,* Auckland, 2-7 July 2005, [CD-ROM, with video segments]. Brisbane, Australia: University of Queensland.

Wild, C.J. (1994). On embracing the ‘wider view’ of statistics. *The American Statistician*, *48*(2), 163-171.

Wild, C. J., and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*(3), 223-265.

CHRIS WILD
Department of Statistics
University of Auckland
Private Bag 92019
Auckland, New Zealand

APPENDIX: USAGE OF DISTRIBUTION

(Below "~" represents "distribution" to focus attention on accompanying adjectives and other modifiers)

NAMED theoretical ~ (e.g., normal, binomial, …)
~ OF SOMETHING OBSERVED
ACTING UPON ~s (e.g., comparing ~s, …)
Some sort of characteristic of ~ (mean, sd, quantiles, …)
Descriptor of (~ is skewed , symmetric, bimodal, long-tailed, …)

vertical ~, horizontal ~, length ~, Circular ~s, spherical ~s, Geographical ~
Of some sort of extreme (e.g., maximum flow ~)
Efficacy ~
Spatial ~, temporal ~
Stationary ~
Spectral ~
Survival ~
Shape ~s
Shot-noise ~s
Latent root ~s
Frequency ~
Empirical ~
Sample ~
Observed ~

Probability ~
Parametric ~
Dependence of ~al shape on parameters
Multiparameter ~s
Continuous versus discrete ~
Derived ~
Probability mass ~
Cumulative ~
Inverse cumulative ~
(cumulative) ~ function
Univariate ~, bivariate ~, multivariate ~
Conditional ~, Marginal ~, Joint ~
Truncated ~
Tolerance ~
Inflated ~s
Run length ~

Target ~
Theoretical ~
Unknown ~
Underlying ~, Population ~, True ~,
Sampling ~
Confidence ~
Null and Alternative ~s (testing)
Nonnull ~ theory
~ of a test statistic
Independence ~

Expected ~
Reference ~
~ free

Permutation ~
bootstrap ~, bootstrap resampling ~
jackknife ~
Simulated ~
Imputation ~
Predictive ~s

Error ~
Residual ~
Studentized ~

~ theory
~al properties
Spaces of ~s
Families of ~s
Asymptotic ~
Limiting ~
Convergence in ~
infinitely divisible ~s

Mixture ~
Mixing ~
Contaminated ~

Initial ~
Equilibrium ~, steady-state ~

Random effect ~
Frailty ~
Latent ~
~ of one or more latent variables

BAYESIAN
Prior ~, hyperprior ~
Posterior ~
Reference prior ~
Improper prior ~
(unmodified, or Bayesian or posterior) predictive ~
(Metropolis-Hasting) candidate ~
Target ~ (in MCMC sampling)
weighted