WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

Design Effect

Tal Galili*

Abstract

In survey research, the design effect is a number that shows how well a sample of people may represent a larger group of people for a specific measure of interest (such as the mean). This is important when the sample comes from a sampling method that is different than just picking people using a simple random sample. The design effect is a positive real number, represented by the symbol Deff. If Deff=1, then the sample was selected in a way that is just as good as if people were picked randomly. When Deff>1, then inference from the data collected is not as accurate as it could have been if people were picked randomly. When researchers use complicated methods to pick their sample, they use the design effect to check and adjust their results. It is also used when planning a study in order to figure out how many people should be in the sample.

Introduction

In survey methodology, the **design effect** (generally denoted as $Deff$, $Deff$, or D^2_{eft}) is a measure of the expected impact of a sampling design on the variance of an estimator for some parameter of a population. It is calculated as the ratio of the variance of an estimator based on a sample from an (often) complex sampling design, to the variance of an alternative estimator based on a simple random sample (SRS) of the same number of elements.[1]:258 The $Deff$ (be it estimated, or known a priori) can be used to evaluate the variance of an estimator in cases where the sample is not drawn using simple random sampling. It may also be useful in sample size calculations[2] and for quantifying the representativeness of samples collected with various sampling designs.

The design effect is a positive real number that indicates an inflation ($Deff > 1$), or deflation ($Deff < 1$) in the variance of an estimator for some parameter, that is due to the study not using SRS (with $Deff = 1$, when the variances are identical).[3]:53,54 Intuitively we can get $Deff < 1$ when we have some a-priori knowledge we can exploit during the sampling process (which is somewhat rare). And, in contrast, we often get $Deff > 1$ when we need to compensate for some limitation in our ability to collect data (which is more common). Some sampling designs that could introduce $Deff$ generally greater than 1 include: cluster sampling (such as when there is correlation between observations), stratified sampling (with disproportionate allocation to the strata sizes), cluster randomized controlled trial, disproportional (unequal probability) sample (e.g. Poisson sampling), statistical adjustments of the data for non-coverage or non-response, and many others. Stratified sampling can yield $Deff$ that is smaller than 1 when using *Proportionate allocation* to strata sizes (when these are known a-priori, and correlated to the outcome of interest) or *Optimum allocation* (when the variance differs between strata and is known a-priori).

Many calculations (and estimators) have been proposed in the literature for how a known sampling design influences the variance of estimators of interest, either increasing or decreasing it. Generally, the design effect varies among different statistics of interests, such as the total or ratio mean. It also matters if the sampling design is correlated with the outcome of interest. For example, a possible sampling design might be such that each element in the sample may have a different probability to be selected. In such cases, the level of correlation between the probability of selection for an element and its measured outcome can have a direct influence on the subsequent design effect. Lastly, the design effect can be influenced by the distribution of the outcome itself. All of these factors should be considered when estimating and using design effect in practice.[4]:13

***Author correspondence**: tal.galili@gmail.com
ORCID: 0000-0003-0046-7332
Supplementary material: *commons.wikimedia.org/location*
Licensed under: CC-BY SA
Received 09-02-2023; accepted 05-05-2024

WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

History

The term "design effect" was coined by Leslie Kish in his 1965 book "Survey Sampling."[1]:88,258 In it, Kish proposed the general definition for the design effect,[a] as well as formulas for the design effect of cluster sampling (with intraclass correlation);[1]:162 and the famous design effect formula for unequal probability sampling.[1]:427 These are often known as "Kish's design effect", and were later combined into a single formula. In a 1995 paper,[5]:73 Kish mentions that a similar concept, termed "Lexis ratio", was described at the end of the 19th century. The closely related Intraclass correlation was described by Fisher in 1950, while computations of ratios of variances were already published by Kish and others from the late 1940s to the 1950s. One of the precursors to Kish's definition was work done by Cornfield in 1951.[6][4]

In his 1995 paper, Kish proposed that considering the design effect is necessary when averaging the same measured quantity from multiple surveys conducted over a period of time.[5]:57–62 He also suggested that the design effect should be considered when extrapolating from the error of simple statistics (e.g. the mean) to more complex ones (e.g. regression coefficients). However, when analyzing data (e.g., using survey data to fit models), Deff values are less useful nowadays due to the availability of specialized software for analyzing survey data. Prior to the development of software that computes standard errors for many types of designs and estimates, analysts would adjust standard errors produced by software that assumed all records in a dataset were i.i.d by multiplying them by a Deft (see *Deft* definition below).

Definitions

Notations

Table 1: Summary of notation

Symbol	Description
$var(\hat{\theta}_w)$	Variance of an estimator $\hat{\theta}_w$ under a given sampling design
$var(\hat{\theta}_{SRSWOR})$	Variance of an estimator $\hat{\theta}_{SRSWOR}$ under simple random sampling without replacement (SRSWOR)
$var(\hat{\theta}_{SRSWR})$	Variance of an estimator $\hat{\theta}_{SRSWR}$ under simple random sampling with replacement (SRSWR)
$Deff$, $Deff$	Design effect, a measure of the impact of a sampling design on the variance of an estimator compared to simple random sampling without replacement (SRSWOR), $Deff_p(\hat{\theta}) = \frac{var(\hat{\theta}_w)}{var(\hat{\theta}_{SRSWOR})}$
$Deft$, $Deft$	Design effect factor, the square root of the ratio of variances under a given sampling design and SRS with replacement (SRSWR), $Deft = \sqrt{\frac{var(\hat{\theta}_w)}{var(\hat{\theta}_{SRSWR})}}$
n	Sample size
N	Population size
n_{eff}	Effective sample size, the sample size under SRS needed to achieve the same variance as the given sampling design, $n_{\text{eff}} = \frac{n}{Deff}$
w_i	Weight for the i-th unit

n_h	Sample size for stratum h
N_h	Population size for stratum h
w_h	Weight for stratum h
H	Total number of strata
$n^*\bar{b}$	Average cluster size
K	Total number of clusters
n_k	Sample size for cluster k
ρ	Intraclass correlation coefficient (ICC) for cluster sampling
$L\, C_{V\ relvar(w)}^2$	Measures of variation in weights using the coefficient of variation (CV) squared (relvariance)
$\hat{\rho}_{y,P}$	Estimated correlation between the outcome variable y and the selection probabilities P
$\hat{\alpha}$	Estimated intercept in the linear regression of the outcome variable y on the selection probabilities P
$\hat{\sigma}_y$	Estimated standard deviation of the outcome variable y
$cv_w\, cv_P$	Coefficient of variation for the weights w and selection probabilities P respectively
f	Sampling fraction, f=n/N
S_y^2	Population variance of the outcome variable y
$p_i\, P_i$	Selection probability for the i-th unit
π_i	Inclusion probability for the i-th unit

Deff

The **design effect,** commonly denoted by $Deff$ (or D_{eff}, sometimes with additional subscripts), is the ratio of two theoretical variances for estimators of some parameter (θ):[1][7]

- The numerator represents the actual variance for an estimator of a parameter ($\hat{\theta}_w$) under a given sampling design p ;
- The denominator represents the variance assuming the same sample size, but if the sample were obtained using the estimator for simple random sampling *without* replacement($\hat{\theta}_{SRSWOR}$).

In other words, $Deff$ measures the extent to which the variance has increased (or, in some

So that:

$$Deff_p(\hat{\theta}) = \frac{var(\hat{\theta}_w)}{var(\hat{\theta}_{SRSWOR})}$$

cases, decreased) because the sample was drawn and adjusted to a specific sampling design (e.g., using weights or other measures) compared to if the sample was from a simple random sample (without replacement). Notice how the definition of $Deff$ is based on parameters of the population that are often unknown, and that are hard to estimate directly. Specifically, the definition involves the variances of estimators under two different sampling designs, even though only a single sampling design is used in practice.

For example, when estimating the population mean, the $Deff$ (for some sampling design p) is:[4]:4[3]:54[b]

$$Deff_p = \frac{var_p(\bar{y}_p)}{(1-f)S_y^2/n}$$

Where n is the sample size, $f = n/N$ is the fraction of the sample from the population, $(1-f)$ is the (squared) finite population correction (FPC), S_y^2 is the unbiassed sample variance, and $var_p(\bar{y}_p)$ is some estimator of the variance of the mean under the sampling design. The issue with the above formula is that it is extremely rare to be able to directly estimate the variance of the estimated mean under two different sampling designs, since most studies rely on only a single sampling design.

There are many ways of calculation $Deff$, depending on the parameter of interest (e.g. population total, population mean, quantiles, ratio of quantities etc.), the estimator used, and the sampling design (e.g. clustered sampling, stratified sampling, post-stratification, multi-stage sampling, etc.).[8]:98 The process of estimating $Deff$ for specific designs will be described in the following section.

Deft

A related quantity to $Deff$, proposed by Kish in 1995, is the *Design Effect Factor*, abbreviated as $Deft$ (or also $Deft$).[5]:56[4] It is defined as the square root of the variance ratios while also having the denominator use a simple random sample *with* replacement (SRSWR), instead of *without replacement* (SRSWOR):

$$Deft = \sqrt{\frac{var(\hat{\theta}_w)}{var(\hat{\theta}_{SRSWR})}}$$

In this later definition (proposed in 1995, vs 1965) Kish argued in favor of using $Deft^2$ over $Deff$ for several reasons. It was argued that SRS "without replacement" (with its positive effect on the variance) should be captured in the denominator part in the definition of the design effect, since it is part of the sampling design. Also, since often the use of the factor is in confidence intervals), it was claimed that using $Deft$ will be simpler than writing $\sqrt{Deff}$. It is also said th at for many cases when the population is very large, $Deft$ is (almost) the square root of $Deff$ ($Deft \approx \sqrt{Deff}$), hence it is easier to use than exactly calculating the finite population correction (FPC). [c]

Even so, in various cases a researcher might approximate the $Deft$ by calculating the variance in the numerator while assuming SRS with replacement (SRSWR) instead of SRS without replacement (SRSWOR), even if it is not precise. For example, consider a multi-stage design with primary sampling units (PSUs) selected systematically with probability proportional to some measure of size from a list sorted in a particular way (say, by number of households in each PSU). Also, let it be combined with an estimator that uses raking to match the totals for several demographic variables. In such a design, the joint selection probabilities for the PSUs, which are needed for a without replacement variance estimator, are o for some pairs of PSUs - implying that an exact design-based (i.e., repeated sampling) variance estimator does not exist. Another example is when a public use file issued by some government agency is used for analysis. In such a case the information on joint selection probabilities of first-stage units is almost never released. As a result, an analyst cannot estimate a with replacement variance for the numerator even if desired. The standard workaround is to compute a variance estimator as if the PSUs were selected with replacement. This is the default choice in software packages such as Stata, the R survey package, and the SAS survey procedures.

WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

Effective sample size

The **effective sample size**, defined by Kish in 1965, is calculated by dividing the original sample size by the design effect.[1]:162,259[9]:190,192 Namely:

$$n_{\text{eff}} = \frac{n}{Deff}$$

This quantity reflects what would be the sample size that is needed to achieve the current variance of the estimator (for some parameter) with the existing design, if the sample design (and its relevant parameter estimator) were based on a simple random sample.[10]

A related quantity is the *effective sample size* ***ratio***, which can be calculated by simply taking the inverse of $Deff$ (i.e., $\frac{n_{eff}}{n} = \frac{1}{Deff}$).

For example, let the design effect, for estimating the population mean based on some sampling design, be 2. If the sample size is 1,000, then the effective sample size will be 500. It means that the variance of the weighted mean based on 1,000 samples will be the same as that of a simple mean based on 500 samples obtained using a simple random sample.

The design effect for well-known sampling designs

The design effect depends on sampling design and statistical adjustments

Different sampling designs and statistical adjustments may have substantially different impact on the bias and variance of estimators (such as the mean).

An example of a design which can lead to estimation efficiency, compared to simple random sampling, is Stratified sampling. This efficiency is gained by leveraging information about the composition of the population. For example, if it is known that gender is correlated with the outcome of interest, and also that the male-female ratio for some population is (say) 50%-50%, then sampling exactly half of the sample from each gender will reduce the variance of the outcome's estimator. Similarly, if a particular sub-population is of special interest, deliberately over-sampling from that sub-population will decrease the variance for estimations made about it.

improvement in variance efficiency might sometimes be sacrificed for convenience or cost. For example, in the cluster sampling case the units may have equal or unequal selection probabilities, irrespective of their intra-class correlation (and their negative effect of increasing the variance of the estimators). We might decide (for practical reasons) to collect responses from only 2 people of each household (i.e., a sampled cluster), which could lead to more complex post-sampling adjustment to deal with unequal selection probabilities. Also, such decisions could lead to less efficient estimators than just taking a fixed proportion of responses from a cluster.

When the sampling design isn't set in advance and needs to be figured out from the data we have, this can lead to an increase both the variance and bias of the weighted estimator. This might happen when making adjustments for issues like non-coverage, non-response, or unexpected strata split of the population that wasn't available during the initial sampling stage. In these cases, we might use statistical procedures such as post-stratification, raking, or inverse propensity score weighting (where the propensity scores are estimated), among other methods. Using these methods requires assumptions about the initial design model. For example, when we use post-stratification based on age and gender, it is assumed that these variables can explain a significant portion of the bias in the sample. The quality of these estimators is closely tied to the quality of the additional information and the missing at random assumptions used when making them. Either way, even when estimators (like propensity score models) do a good job capturing most of the sampling design, using the weights can make a small or a large difference, depending on the specific data-set.

Due to the large variety in sampling designs (with or without an effect on unequal selection probabilities), different formulas have

been developed to capture the potential design effect, as well as to estimate the variance of estimators when accounting for the sampling designs.[11] Sometimes, these different design effects can be compounded together (as in the case of unequal selection probability and cluster sampling, more details in the following sections). Whether or not to use these formulas, or just assume SRS, depends on the expected amount of bias reduction vs. the increase in estimator variance (and in the overhead of methodological and technical complexity).[1]:426

Table 2: Summary of design effect formulas

Formula Name	Equation	Description
Kish's design effect for unequal weights	$Deff = \frac{n\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2} = \frac{\overline{w^2}}{\overline{w}^2}$	Measures the loss in precision due to unequal weights, where w_i is the weight for the i-th unit.
Kish's design effect for cluster sampling	$Deff = 1 + (n^* - 1)\rho$	Measures the loss in precision due to cluster sampling, where n^* is the average cluster size and ρ is the intraclass correlation.
Kish's combined design effect	$Deff = \frac{n\sum_{h=1}^{H}(n_h w_h^2)}{(\sum_{h=1}^{H} n_h w_h)^2}(1 + (n^* - 1)\rho)$	Measures the combined effect of unequal weights and cluster sampling, where n_h and w_h are the sample size and weight for the h-th stratum, respectively.
Spencer's design effect for estimated total	$Deff = (1 - \hat{\rho}_{y,P}^2)(1 + L) + (\frac{\hat{\alpha}}{\hat{\sigma}_y})^2 L$	Measures the design effect for estimating a total when there is a correlation between the outcome and the selection probabilities, where $\hat{\rho}_{y,P}$ is the estimated correlation, L is the relvariance of the weights, $\hat{\alpha}$ is the estimated intercept, and $\hat{\sigma}_y$ is the estimated standard deviation of the outcome.
Park and Lee's design effect for estimated ratio mean	$Deff = (1 - \hat{\rho}_{y,P}^2)(1 + cv_w^2) + \frac{\hat{\rho}_{y,P}^2}{cv_P^2} cv_w^2$	Measures the design effect for estimating a ratio mean when there is a correlation between the outcome and the selection probabilities, where cv_w and cv_P are the coefficients of variation for the weights and selection probabilities, respectively.
Henry's design effect for calibration weighting	Extends Kish's design effect to include calibration weighting in single-stage samples	Proposes a model-assisted design effect measure for single-stage sampling with calibration weighting, considering the correlation between the outcome and the calibration variables.
Lohr's design effect for regression slope	Provides design effect formulas for OLS and GLS regression slope estimators in cluster sampling	Presents design effect formulas for ordinary least squares (OLS) and generalized least squares (GLS) regression slope estimators in the context of cluster sampling, using a random coefficient model.

WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

Unequal selection probabilities

Sources of unequal selection probabilities

Table 3: Summary of sources for unequal selection probabilities

Source	Description	Examples	Impact on Sampling Probabilities
Disproportional sampling	Deliberately over/under sampling specific sub-populations or clusters	- Optimum allocation in stratified sampling - Oversampling smaller groups for comparison - Cluster sampling with unequal cluster sizes	Leads to unequal selection probabilities by design
Non-coverage	Failure to include all elements of the target population in the sampling frame	- Sampling based on incomplete lists (e.g., phone books) - Advertising to recruit survey participants	Affects sampling probabilities, but the impact is difficult to measure and adjust for
Non-response	Failure to obtain measurements from sampled units that were intended to be measured	- Unit non-response (e.g., refusal, not-at-home) - Item non-response (e.g., sensitive questions) - Inability to respond (e.g., language barrier, illness)	Leads to unequal selection probabilities, as non-response rates may vary across subgroups
Statistical adjustments	Post-hoc adjustments to the sample weights to account for known population characteristics or to mitigate non-coverage and non-response biases	- Post-stratification - Raking - Propensity score weighting - Calibration weighting	Introduces unequal weights to improve representativeness, but may increase variance

There are various ways to sample units so that each unit would have the exact same probability of selection. Such methods are called equal probability sampling (EPSEM) methods. Some of the more basic methods include simple random sampling (SRS, with or without replacement) and systematic sampling for getting a fixed sample size. There is also Bernoulli sampling with a random sample size. More advanced techniques such as stratified sampling and cluster sampling can also be designed to be EPSEM. For example, in cluster sampling we can use a two stage sampling in which we sample each cluster (which may be of different sizes) with equal probability, and then sample from each cluster at the second stage using SRS with a fixed proportion (e.g. sample half of the cluster, the whole cluster, etc.). This method will yield EPSEM, but the specific number of elements we end up with is stochastic (i.e., non deterministic).[d][12]:3–8 Another strategy for cluster sampling that leads to EPSEM is to sample clusters in a way that is proportional to their sizes, and then sample a fixed number of elements inside each cluster.[e]

In their works, Kish and others highlights several known reasons that lead to unequal selection probabilities:[1]:425[9]:185[5]:69[13]:50,395[14]:306

1. **Disproportional sampling** due to selection frame or procedure. This happens when a researcher deliberately over- or under-samples specific sub-populations or clusters. For example:

a) In stratified sampling when units from some strata are known to have a larger variance than other

strata. In such cases, the intention of the researcher may be to use this prior knowledge about the variance between strata in order to reduce the overall variance of an estimator of some population level parameter of interest (e.g., the mean). This can be achieved by a strategy known as *optimum allocation*, in which a stratum h is over sampled proportional to higher standard deviation and lower sampling cost (i.e., $f_h \propto \frac{S_h}{\sqrt{C_h}}$, where S_h is the standard deviation of the outcome in h, and C_h relates to the cost of recruiting one element from h). An example of an optimum allocation is *Neyman's optimal allocation* which, when cost is fixed for recruiting people from each stratum, the sample size is: $n_h = n\frac{W_h S_{Uh}}{\sum_h W_h S_{Uh}}$. Where the summation is over all strata: *n* is the total sample size; n_h is the sample size for stratum *h*; $W_h = \frac{N_h}{N}$ is the relative size of stratum *h* as compared to the entire population *N*; and S_{Uh} is the standard error in stratum *h*.[15] A related concept to optimum design is optimal experimental design.

b) If there is interest in comparing two strata (e.g., people from two specific socio-demographic groups, or from two regions, etc.), in which case the smaller group may be over-sampled. This way, the variance of the estimator that compares the two groups is reduced.

c) In cluster sampling there may be clusters of different sizes but the procedure samples from all clusters using SRS, and all elements in the cluster are measured (for example, if the cluster sizes are not known upfront at the stage of sampling).

d) In some two-stage cluster sampling based cluster sizes. For example, when in the first stage the clusters are sampled proportionally to the estimation of their size (a.k.a.: **PPS** Probability Proportional to Size) and at the second stage a fixed proportion of elements are chosen (e.g., half, or all the elements in the cluster) - then the selection probabilities are different for elements from different clusters. A similar case is when the first stage attempts to sample the clusters using PPS, the second stage uses a fixed number of elements in each cluster - but the cluster sizes used for the first stage sampling were inaccurate (so that some smaller cluster may have a higher-than-it-should chance of being selected. And vice versa for larger clusters with too-small a chance of being sampled). In such cases, the larger the errors in the sampling probabilities used in the first stage, the larger the unequal selection probabilities for each element will be.[8]:109[f]

e) When the frame used for sampling includes duplication of some of the items, thus leading some items to have a larger probability than others to be sampled (e.g., if the sampling frame was created by merging several lists. Or if recruiting users from several ad channels in which some of the users are available for recruitment from several of the channels, while others are available to be recruited from only one of the channels) so that different units would have different sampling probabilities, thus making this sampling procedure to not be EPSEM.[12]:3–8[9]:186

f) When several different samples/frames are to be combined. For example, if running different ad campaigns for recruiting respondents. Or when combining results from several studies done by different researchers and/or at different times (i.e., Meta-analysis).[9]:188

When disproportional sampling happens, due to sampling design decisions, the researcher may (sometimes) be able to trace back the decision and accurately calculate the exact inclusion probability. When these selection probabilities are hard to trace back, they may be estimated using some propensity score model combined with information from auxiliary variables (e.g., age, gender, etc.).

2. **Non-coverage**.[1]:527,528 This happens, for example, if people are sampled based on some pre-defined list that doesn't include all the people in the population (e.g., a phone book or using ads to recruit people to a survey). These missing units are missing due to some failure of creating the sampling frame, as opposed to deliberate exclusion of some people (e.g. minors, people who cannot vote, etc.). The effect of non-coverage on sampling probability is considered difficult to measure (and adjust for) in various survey situations, unless strong assumptions are made. Adjustments for non-coverage can lead to inadequate weights when the relevant covariates are not used for adjustment. If there are covariates that can be used to correct for non-coverage, they are expected to lead to unequal survey weights.

3. **Non-response**. This refers to the failure of obtaining measurements on sampled units that are intended to be measured. Reasons for non-response are varied and depends on the context. A person may be temporarily unavailable, for example if they are not available to answer the phone when a telephone survey is done. A person may also refuse to answer the survey due to a variety of reasons, e.g. different tendencies of people from different ethnic/demographic/socio-economic groups to respond in general; insufficient incentive to spend the time or share data; the identity of the institution that is running the survey; inability to respond (e.g. due to illness, illiteracy, or a language barrier); respondent is not found (e.g. they moved); the response was lost/destroyed during encoding or

transmission (i.e., measurement error). In the context of surveys, these reasons may be related to answering the entire survey or just specific questions.[1]:532[9]:186

4. **Statistical adjustments**. These may include methods such as post-stratification, raking, or propensity score (estimation) models - used to perform an adjustment of the sample to some known (or estimated) strata sizes. These adjustments can be in addition of *design weights*, which aims to account for imbalances due to some known sampling design. Such procedures are used to mitigate issues in the sampling ranging from sampling error, under-coverage of the sampling frame to non-response.[16]:45[17] For example, these methods can be used to make the sample more similar to some target "controls" (i.e., population of interest), a process also called "standardization".[9]:187 In such cases, these adjustments help with providing unbiased estimators (often with the cost of increased variance, as seen in the following sections). If the original sample is a nonprobability sample, then post-stratification adjustments are just similar to quota sampling.[9]:188,189. Note that if a simple random sample is used, a post-stratification (using some auxiliary information) does not offer an estimator that is uniformly better than just an unweighted estimator. However, it can be viewed as a more "robust" estimator.[18] Alternatively, when the sampling design is fully known (leading to some p_h probability of selection for some element from stratum *h*), and the non-response is measurable (i.e., we know that only r_h observations answered in stratum *h*), then an exactly known inverse probability weight can be calculated for each element *i* from stratum *h* using: $w_i = \frac{1}{p_h r_h}$.[9]:186[g] Sometimes a statistical adjustment, such as post-stratification or raking, is used for estimating the selection probability. E.g., when comparing the sample we have with same target population, also known as matching to controls. The estimation process may be focused only on adjusting the existing population to an alternative population (for example, if trying to extrapolate from a panel drawn from several regions to an entire country). In such a case, the adjustment might be focused on some calibration factor c_i and the weights be calculated as $w_i = \frac{c_i}{p_h r_h}$ [9]:187 However, in other cases, both the under-coverage and non-response are all modeled as part of the statistical adjustment, which leads to an estimation of the overall sampling probability (lets say p'_i). In such a case, the weights are simply: $w_i = \frac{1}{p'_i}$. Notice that when statistical adjustments are used, w_i is often estimated based on some model. The formulation in the following sections assume this w_i is known, which is not true for statistical adjustments (since we only have $\widehat{w}_i$). However, if it is assumed that the estimation error of $\widehat{w}_i$ is very small then the following sections can be used as if it was known. Having this assumption be true depends on the size of the sample used for modeling, and is worth keeping in mind during analysis. When the selection probabilities may be different, the sample size is random, and the pairwise selection probabilities are independent, we call this Poisson sampling.[19]

"Design based" vs "model based" for describing properties of estimators

Adjusting for unequal probability selection through "individual case weights" (e.g. inverse probability weighting), yields various types of estimators for quantities of interest. Estimators such as Horvitz–Thompson estimator yield unbiased estimators (if the selection probabilities are indeed known, or approximately known), for total and the mean of the population. Deville and Särndal (1992) coined the term "**calibration estimator**" for estimators using weights such that they satisfy some condition, such as having the sum of weights equal the population size. And more generally, that the weighted sum of weights is equal some quantity of an auxiliary variable: $\sum w_i x_i = X$(e.g., that the sum of weighted ages of the respondents is equal to the population size in each age group).[20][17]:132[21]:1

The two primary ways to argue about the properties of calibration estimators are:[17]:133–134[22]

1. **randomization based** (or, sampling design based) - in this case, the weights (w_i) and values of the outcome of interest y_i that are measured in the sample are all treated as known. In this framework, there is variability in the (known) values of the outcome (**Y**). However, the only randomness comes from which of the elements in the population were picked into the sample (often denoted as I_i , getting 1 if element i is in the sample and 0 if it is not). For a simple random sample, each I_i will be an IID Bernoulli distribution with some parameter p . For general EPSEM (equal probability sampling) I_i will still be Bernoulli with some parameter p, but they may no longer be independent random variables. I.e., knowing that a sample is EPSEM means that it maintains marginally equal probability of selection, but it does

not inform us about the joint probability of selection. For something like post stratification, the number of elements at each stratum can be modeled as a multinomial distribution with different p_h inclusion probabilities for each element belonging to some stratum h. In these cases, the sample size itself can be a random variable.

2. **model based** - in this case, the sample is fixed, the weights are fixed, but the outcome of interest is treated as a random variable. For example, in the case of post-stratification, the outcome can be modeled as some linear regression function where the independent variables are indicator variables mapping each observation to its relevant stratum, and the variability comes with the error term.

As we will see later, some proofs in the literature rely on the randomization-based framework, while others focus on the model-based perspective. When moving from the mean to the weighted mean, more complexity is added. For example, in the context of survey methodology, often the population size itself is considered an unknown quantity that is estimated. So in the calculation of the weighted mean is in fact based on a ratio estimator, with an estimator of the total at the numerator and an estimator of the population size in the denominator (making the variance calculation to be more complex).[23]

Common types of weights

Table 4: Summary of common types of weights used in design effect calculations

Weight Type	Description	Interpretation
Frequency weights	Each weight is an integer indicating the absolute frequency of an item in the sample	Specific value has an absolute meaning; weights represent the amount of information in the dataset
Inverse-variance weights	Each element is assigned a weight that is the inverse of its known variance	When all elements have the same expectancy, using such weights for weighted averages has the least variance
Normalized (convex) weights	Weights form a convex combination (sum to 1); can be normalized to sum to sample size (n)	Weights that sum to n have a relative interpretation: elements with weights > 1 are more "rare" than average and have larger influential on (say) the average, while weights < 1 are more "common" and less influential
Inverse probability weights	Each element is given a weight proportional to the inverse of its selection probability	Weights represent how many items each element "represents" in the target population; sum of weights equals the size of the target population

There are many types (and subtypes) of weights, with different ways to use and interpret them. With some weights their absolute value has some important meaning, while with other weights the important part is the relative values of the weights to each other. This section introduces some of the more common types of weights so that they can be referenced in follow-up sections.

- **Frequency weights**[24] are a basic type of weighting presented in introductory statistics courses. With these, each weight is an integer number that indicates the absolute frequency of an item in the sample. These are also sometimes termed repeat (or occurrence) weights. The specific value has an absolute meaning that is lost if the weights are transformed, such as when scaling. For example: if we have the numbers 10 and 20 with the frequency weights values of 2 and 3, then when "spreading" our data it is: 10,10, 20, 20, 20 (with weights of 1 to each of these items). Frequency weights includes the amount of information contained in a dataset, and thus allows things like creating unbiased weighted variance estimation using Bessel's correction. No-

tice that such weights are often random variables, since the specific number of items we will see from each value in the dataset is random.

- **inverse-variance weighting**, also known as *analytic weights*[24], is when each element is assigned a weight that is the inverse of its (known) variance.[25][9]:187 When all elements have the same expectancy, using such weights for calculating weighted averages has the least variance among all weighted averages. In the common formulation, these weights are known and not random.

- **Normalized (convex) weights** is a set of weights that form a convex combination, i.e., each weight is a number between 0 and 1, and the sum of all weights is equal to 1. Any set of (non negative) weights can be turned into normalized weights by dividing each weight with the sum of all weights, making these weights normalized to sum to 1.

A related form are **weights normalized to sum to sample size (n)**. These (non-negative) weights sum to the sample size (n), and their mean is 1. Any set of weights can be normalized to sample size by dividing each weight with the average of all weights. These weights have a nice relative interpretation where elements with weights larger than 1 are more "influential" (in terms of their relative influence on, say, the weighted mean) then the average observation, while weights smaller than 1 are less "influential" than the average observation.

- **Inverse probability weighting**, or simply *probability weights*[24], is when each element is given a weight that is (proportional) to the inverse probability of selecting that element. E.g., by using $w_i = \frac{1}{p_i}$.[9]:185 With inverse probability weights, we learn how many items each element "represents" in the target population. Hence, the sum of such weights returns the size of the target population of interest. Inverse probability weights can be normalized to sum to 1 or normalized to sum to the sample size (n), and many of the calculations from the following sections will yield the same results.

When a sample is EPSEM then all the probabilities are equal and the inverse of the selection probability yield weights that are all equal to one another (they are all equal to $\frac{N}{n} = \frac{1}{f}$, where n is the sample size and N is the population size). Such a sample is called a **self weighting sample**.[9]:193

There are also indirect ways of applying "weighted" adjustments. For example, the existing cases may be duplicated to impute missing observations (e.g. from non-response), with variance estimated using methods such as multiple imputation. An alternative approach is to remove (assign a weight of 0 to) some cases. For example, when wanting to reduce the influence of over-sampled groups that are less essential for some analysis. Both cases are similar in nature to inverse probability weighting but the application in practice gives more/less rows of data (making the input potentially simpler to use in some software implementation), instead of applying an extra column of weights. Nevertheless, the consequences of such implementations are similar to just using weights. So while in the case of removing observations the data can easily be handled by common software implementations, the case of adding rows requires special adjustments for the uncertainty estimations. Not doing so may lead to erroneous conclusions(i.e., there is no free lunch when using alternative representation of the underlying issues).[9]:189,190

The term "Haphazard weights", coined by Kish, is used to refer to weights that correspond to unequal selection probabilities, but ones that are not related to the expectancy or variance of the selected elements.[9]:190,191

WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

Haphazard weights with estimated ratio-mean ($\hat{\bar{Y}}$) - Kish's design effect

Formula

When taking an unrestricted sample of n elements, we can then randomly split these H elements into disjoint strata, each of them containing some size of n_h elements so that $\sum_{h=1}^{H} n_h = n$. All elements in each stratum h has some (known) non-negative weight assigned to them (w_h). The weight w_h can be produced by the inverse of some unequal selection probability for elements in each stratum h (i.e., inverse probability weighting following a procedure such as post-stratification). In this setting, **Kish's design effect**, for the increase in variance **of the sample weighted mean** due to this design (reflected in the weights), versus SRS of some outcome variable y (when there is no correlation between the weights and the outcome, i.e. haphazard weights) is:[1]:427[9]:191(4.2)

$$Deff = \frac{n \sum_{h=1}^{H} (n_h w_h^2)}{(\sum_{h=1}^{H} n_h w_h)^2}$$

By treating each item as coming from its own stratum $\forall h : n_h = 1$, Kish (in 1992) simplified the above formula to the (well-known) following version:[9]:191(4.3)[26]:318[4]:8

$$Deff = \frac{n \sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2} = \frac{\frac{1}{n} \sum_{i=1}^{n} w_i^2}{\left(\frac{1}{n} \sum_{i=1}^{n} w_i\right)^2} = \frac{\overline{w^2}}{\overline{w}^2}$$

This version of the formula is valid when one stratum had several observations taken from it (i.e., each having the same weight), or when there are just many strata were each one had one observation taken from it, but several of them had the same probability of selection. While the interpretation is slightly different, the calculation of the two scenarios comes out to be the same.

When using Kish's design effect for unequal weights, you may use the following simplified formula for "Kish's Effective Sample Size"[27][1]:162,259

$$n_{\text{eff}} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

Proof **Expand**

Assumptions and Proofs

The above formula, by Kish, gives the increase in the variance of the weighted mean based on "haphazard" weights. This can also be written as the following formula where y are observations selected using unequal selection probabilities (with no within-cluster correlation, and **no relationship** to the expectancy or variance of the outcome measurement),[9]:190,191 and y' are the observations we would have had if we got them from a simple random sample:

$$Deff_{kish} = \frac{var\left(\bar{y}_w\right)}{var\left(\bar{y}'\right)} = \frac{var\left(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}\right)}{var\left(\frac{\sum_{i=1}^{n} y'_i}{n}\right)}$$

It can be shown that the ratio of variances formula can be reduced to Kish's formula by using a model based perspective.[28] In it, Kish's formula will hold when all n observations ($y_1, \ldots, y_n$) are (at least approximately) uncorrelated ($\forall (i \neq j) : cor(y_i, y_j) = 0$), with the same variance (σ^2) in the response variable of interest (y). It will also be required to assume the weights themselves are not a random variable but rather some known constants (e.g. the inverse of probability of selection, for some pre-determined and known sampling design).

Proof	Expand

The conditions on y are trivially held if the y observations are IID with the same expectation and variance. In such cases, $y = y'$, and we can estimate $var\left(\bar{y}_w\right)$ by using $\widehat{var\left(\bar{y}_w\right)} = \widehat{var\left(\bar{y}\right)} \times Deff$.[9][29] If the y's are not all with the same expectations then we cannot use the estimated variance for calculation, since that estimation assumes that all y_is have the same expectation. Specifically, if there is a correlation between the weights and the outcome variable y, then it means that the expectation of y is not the same for all observations (but rather, dependent on the specific weight value for each observation). In such a case, while the design effect formula might still be correct (if the other conditions are met), it would require a different estimator for the variance of the weighted mean. For example, it might be better to use a weighted variance estimator.

If different y_is values have different variances, then while the weighted variance could capture the correct population-level variance, Kish's formula for the design effect may no longer be true.

A similar issue happens if there is some correlation structure in the samples (such as when using cluster sampling).

Relation to the coefficient of variation

Notice that Kish's definition of the design effect is closely tied to the coefficient of variation (Kish also calls it *relvariance* or *relvar* for short[h]) of the weights (when using the uncorrected (population level) sample standard deviation for estimation). This has several notations in the literature:[9]:191[13]:396

$$Deff = 1 + L = 1 + {C_V}^2 = 1 + relvar(w) = 1 + \frac{V(w)}{\bar{w}^2}$$

Where $V(w) = \frac{\sum(w_i - \bar{w})^2}{n}$ is the population variance of w, and $\bar{w} = \frac{\sum w_i}{n}$ is the mean. When the weights are normalized to sample size (so that their sum is equal to n and their mean is equal to 1), then ${C_V}^2 = V(w)$ and the formula reduces to $Deff = 1 + V(w)$. While it is true we assume the weights are fixed, we can think of their variance as the variance of an empirical distribution defined by sampling (with equal probability) one weight from our set of weights (similar to how we would think about the correlation of x and y in a simple linear regression).

WikiJournal of Science, 2024, 7(1):4
doi: 10.15347/wjs/2024.004
Encyclopedic Review Article

Proof	Expand

Relation to disproportionate stratified sampling

Kish's original definition compared the variance under some sampling design to the variance achieved through a simple random sample. Some literature provide the following alternative definition for Kish's design effect: "the ratio of the variance of the weighted survey mean under disproportionate stratified sampling to the variance under proportionate stratified sampling when all stratum unit variances are equal".[26]:318[13]:396 Reflecting on this, Park and Lee (2006) stated that "The rationale behind [...][Kish's] derivation is that the loss in precision of [the weighted mean] due to haphazard unequal weighting can be **approximated** by the ratio of the variance under disproportionate stratified sampling to that under the proportionate stratified sampling".[4]:8

Note that this alternative definition only approximated since if the denominator is based on "proportionate stratified sampling" (achieved via stratified sampling) then such a selection will yield a reduced variance as compared with simple random sample. This is since stratified sampling removes some of the variability in the specific number of elements per stratum, as occurs under SRS.

Relatedly, Cochran (1977) provides a formula for the proportional increase in variance due to deviation from optimum allocation (what, in Kish's formulas, would be called *L*).[3]:116

Alternative naming conventions

Early papers used the term $Deff$.[9]:192 As more definitions of the design effect appeared, Kish's design effect for unequal selection probabilities was denoted $Deff_{kish}$ (or $Deft^2_{kish}$) or simply $deff_K$ for short.[4]:8[13]:396[26]:318 Kish's design effect is also known as the "Unequal Weighting Effect" (or just UWE), termed by Liu et al. in 2002.[30]:2124

When the outcome correlates with the selection probabilities

Spencer's Deff for estimated total (Y^)

The estimator for the total is the "p-expanded with replacement" estimator (a.k.a.: *pwr-estimator* or Hansen and Hurwitz). It is based on a simple random sample (with replacement, denoted *SIR*) of *n* items (y_k) from a population of size N.[i] Each item has a probability of p_k (k from 1 to N) to be drawn in a single draw ($\sum_U p_k = 1$, i.e. it is a multinomial distribution). The probability that a specific y_k will appear in the sample is p_k. The "p-expanded with replacement" value is $Z_i = \frac{y_k}{p_k}$ with the following expectancy: $E[Z_i] = E[I_i \frac{y_k}{p_k}] = \frac{y_k}{p_k} E[I_i] = \frac{y_k}{p_k} p_k = y_k$. Hence $\hat{Y}_{pwr} = \frac{1}{n} \sum_i^n Z_i$, the pwr-estimator, is an unbiased estimator for the sum total of y.[3]:51

In 2000, Bruce D. Spencer proposed a formula for estimating the design effect for the variance of estimating the **total** (not the mean) of some quantity ($\hat{Y}$), when there is correlation between the selection probabilities of the elements and the outcome variable of interest.[31]

In this setup, a sample of size *n* is drawn (with replacement) from a population of size *N*. Each item is drawn with probability P_i (where $\sum_{i=1}^{N} P_i = 1$, i.e. multinomial distribution). The selection probabilities are used to define the Normalized (convex) weights: $w_i = \frac{1}{nP_i}$. Notice that for some random set of *n* items, the sum of weights will be equal to 1 only by expectation ($E[w_i] = 1$) with some variability of the sum around it (i.e., the sum of elements from a Poisson binomial distribution). The relationship between y_i and P_i is defined by the following (population) simple linear regression: $y_i = \alpha + \beta P_i + \epsilon_i$

Where y_i is the outcome of element *i*, which linearly depends on P_i with the intercept α and slope β. The residual from the fitted line is $\epsilon_i = y_i - (\alpha + \beta P_i)$. We can also define the population variances of the outcome and the residuals as σ_y^2 and σ_ϵ^2. The correlation between P_i and y_i is $\rho_{y,P}$.

Spencer's (approximate) design effect for estimating the total of *y* is:[31]:138[32]:4[13]:401

$$Deff_{Spencer} = (1 - \hat{\rho}_{y,P}^2)(1 + L) + \left(\frac{\hat{\alpha}}{\hat{\sigma}_y}\right)^2 L$$

Where:

- $\hat{\rho}_{y,P}^2$ estimates $\rho_{y,P}^2$
- $\hat{\alpha}$ estimates the slope α
- $\hat{\sigma}_y$ estimates the population variance σ_y, and
- L is the *relvariance* of the weights, as defined in Kish's formula: $L = cv_w^2 = relvar(w) = \frac{V(w)}{\bar{w}^2}$.

This assumes that the regression model fits well so that the probability of selection and the residuals are independent, since it leads to the residuals, and the square residuals, to be uncorrelated with the weights, i.e., that $\rho_{\epsilon,W} = 0$ and also $\rho_{\epsilon^2,W} = 0$.[31]:138

When the population size (N) is very large, the formula can be written as:[26]:319

$$Deff_{Spencer} = (1 - \hat{\rho}_{y,P}^2)(1 + cv_w^2) + \left(\frac{1}{cv_Y^2}\right)^2 cv_w^2$$

(since $\alpha = \bar{Y} - \beta \times \bar{P} = \bar{Y} - \beta \times \frac{1}{N} \approx \bar{Y}$, where $cv_Y^2 = \frac{\sigma_Y^2}{\bar{Y}^2}$)

This approximation assumes that the linear relationship between *P* and *y* holds. And also that the correlation of the weights with the errors, and the errors squared, are both zero. I.e., $\rho_{w,e} = 0$ and $\rho_{w,e^2} = 0$.[32]:4

We notice that if $\hat{\rho}_{y,P} \approx 0$, then $\hat{\alpha} \approx \bar{y}$ (i.e., the average of *y*). In such a case, the formula reduces to

$$Deff_{Spencer} = (1 + L) + \left(\frac{1}{relvar(y)}\right)^2 L$$

Only if the variance of *y* is much larger than its mean, then the right-most term is close to 0 (i.e., $\frac{1}{relvar(y)} = \frac{\bar{Y}}{\sigma_y} \approx 0$), which reduces Spencer's design effect (for the estimated total) to be equal to Kish's design effect (for the ratio means):[32]:5 $Deff_{Spencer} \approx (1 + L) = Deff_{Kish}$. Otherwise, the two formulas will yield different results, which demonstrates the difference between the design effect of the total vs. the design effect of the mean.

Park and Lee's Deff for estimated ratio-mean (Y¯^)

In 2001, Park and Lee extended Spencer's formula to the case of the ratio-mean (i.e., estimating the mean by dividing the estimator of the total with the estimator of the population size). It is:[32]:4

$$Deff_{Park\&Lee} = (1 - \hat{\rho}_{y,P}^2)(1 + cv_w^2) + \frac{\hat{\rho}_{y,P}^2}{cv_P^2} cv_w^2$$

Where:

- cv_P^2 is the (estimated) squared coefficient of variation of the probabilities of selection.

Park and Lee's formula is exactly equal to Kish's formula when $\hat{\rho}_{y,P}^2 = 0$. Both formulas relate to the design effect of the mean of *y*, while Spencer's $Deff$ relates to the estimation of the population total.

In general, the $Deff$ for the total ($\hat{Y}$) tends to be less efficient than the $Deff$ for the ratio mean ($\hat{\bar{Y}}$)

when $\rho_{y,P}$ is small. And in general, $\rho_{y,P}$ impacts the efficiency of both design effects.[4]:8

Cluster sampling

For data collected using cluster sampling we assume the following structure:

- n_k observations in each cluster and *K* clusters, and with a total of $n = \sum n_k$ observations.
- The observations have a block diagonal correlation matrix in which every pair of observations from the same cluster is correlated with an intra-class correlation of ρ, while every pair from difference clusters are uncorrelated.[33] I.e., for every pair of observations, i and j, if they belong to the same cluster k, we get $cov(y_i, y_j) = \rho\sigma^2$. And two items from two different clusters are not correlated, i.e.: $cov(y_i, y_j) = 0$.
- An element from any cluster is assumed to have the same variance: $var(y_i) = \sigma_h^2 = \sigma^2$.

When clusters are all of the same size n^*, the design effect D_{eff}, proposed by Kish in 1965 (and later re-visited by others), is given by:[1]:162[13]:399[4]:9[34][35][14]:241

$$Deff = 1 + (n^* - 1)\rho.$$

It is sometimes also denoted as $Deff_C$.[30]:2124

In various papers, when cluster sizes are not equal, the above formula is also used with n^* as the average cluster size (which is also sometimes denoted as $\bar{b}$).[36][28]:105 In such cases, Kish's formula (using the average cluster weight) serves as a conservative (upper bound) of the exact design effect.[28]:106

Alternative formulas exists for unequal cluster sizes.[1]:193 Follow up work had discussed the sensitivity of using the average cluster size with various assumptions.[37]

The design effect for complex designs

Unequal selection probabilities × Cluster sampling

In a 1987 paper, Kish proposed a combined design effect that incorporates both the effects due to weighting that accounts for unequal selection probabilities and cluster sampling:[36]:16[28]:105[38]:4[32]:2

$$Deff_{Kish} = \frac{n\sum_{h=1}^{H}(n_h w_h^2)}{\left(\sum_{h=1}^{H} n_h w_h\right)^2}(1 + (n^* - 1)\rho) = deff_k \times deff_C$$

The above uses notations similar to what is used in this article (the original 1987 publication used different notation).[j] A model based justification for this formula was provided by Gabler et al.[28]

Stratified sampling × unequal selection probabilities × Cluster sampling

In 2000, Liu and Aragon proposed a decomposition of unequal selection probabilities design effect for different strata in stratified sampling.[39] In 2002, Liu et al. extended that work to account for stratified samples, where within each stratum is a set of unequal selection probability weights. The cluster sampling is either global or per stratum.[30] Similar work was done also by Park et al. in 2003.[40]

Chen-Rust $Deff$ Deff: Design effects to two- and three-stage designs with stratification

The Chen-Rust $Deff$ extends the model-based justification of Kish's 1987 formula for design effects proposed by Gabler, el. al. [28], applying it to two-stage designs with stratification at the first stage and to three-stage designs without stratification.[41] The modified formulae define the overall design effect using survey weights and population intracluster correlations. These formulae allow for insightful interpretations of design effects from various sources and can estimate intracluster correlations in completed surveys or predict design effects in future surveys.

Henry's $Deff$ Deff: a design effect measure for calibration weighting in single-stage samples

Henry's $Deff$[26] proposes an extended model-assisted weighting design-effect measure for single-stage sampling and calibration weight adjustments for a case where $y_i = \alpha + \beta x_i + \epsilon_i$, where x_i is a vector of covariates, the model errors are independent, and the estimator of the population total is the general regression estimator (GREG) of Särndal, Swensson, and Wretman (1992).[3] The new measure considers the combined effects of non-epsem sampling design, unequal weights from calibration adjustments, and the correlation between an analysis variable and the auxiliaries used in calibration.

Lohr's $Deff$ Deff: a design effect for a regression slope in a cluster sample

Lohr's $Deff$[42] is for ordinary least squares (OLS) and generalized least squares (GLS) estimators in the context of cluster sampling, using a random coefficient regression model. Lohr presents conditions under which the GLS estimator of the regression slope has a design effect less than 1, indicating higher efficiency. However, the design effect of the GLS estimator is highly sensitive to model specification. If an underlying random coefficient model is incorrectly specified as a random intercept model, the design effect can be seriously understated. In contrast, the OLS estimator of the regression slope and the design effect calculated from a design-based perspective are robust to misspecification of the variance structure, making them more reliable in situations where the model specification may not be accurate.

Uses

$Deff$ may be used when planning a future data collection, as well as a diagnostic tool:[14]:85

- **When planning a future data collection** - $Deff$ may be used to evaluate the sampling efficiency. E.g. if there is potentially "too much" increase in variance due to some sampling design decision, or if some alternative (economically feasible) design is more efficient. This also influences the sample size (overall, per stratum, per cluster, etc.). When planning the sample size, work may be done to correct the design effect so as to separate the interviewer effect (measurement error) from the effects of the sampling design on the sampling variance.[43]

- **As a diagnostic tool** - $Deff$ may help in evaluating potential problems with a post-hoc weighting analysis (e.g. from non-response adjustments).[8] For example, if the $Deff$ value is especially high, then it might indicate an issue with the sampling or weighting scheme. This can also assist when performing some manipulation on the weights (e.g., weight trimming), the design effect could be used to evaluate the influence of the manipulation on the effective sample size.[44] And also in identifying glaring issues with the data or its analysis (e.g., ranging from mistakes to the presence of Outliers).[9]:191 Although some literature suggests that $Deff > 1.5$ is likely to require some attention,[13]:396 there is no universal rule of thumb for which design effect value is "too high". Practical considerations of $Deff$ values are often context dependent.

Considering the design effect is unnecessary when[5]:57–62 the source population is closely IID, or when the sample design of the data was drawn as a simple random sample. It is also less useful when the sample size is relatively small (at least partially, for practical reasons).

While Kish originally hoped to have the design effect be as agnostic as possible to the underlying distribution of the data, sampling probabilities, their correlations, and the statistics of interest, follow up research has shown that these do influence the design effect. Hence, these properties should be carefully considered when deciding which $Deff$ calculation to use, and how to use it.[4]:13[32]:6

The design effect is rarely applied when constructing confidence intervals. Ideally, one would be able to determine, for an estimator of a particular parameter, both the variance under Simple Random Sample (SRS) with replacement and the design effect (which accounts for all elements of the sampling design that change the variance). In such scenarios, the basic variance and the design effect could have been multiplied to compute the variance of the estimator for the specific design.[1]:259 This computed value can then be employed to form confidence intervals. However, in real-world applications, it is uncommon to estimate both values simultaneously. As a result, other methods are favored. For instance, Taylor linearization is utilized to construct confidence intervals based on the variance of the weighted mean. More broadly, the bootstrap method, also known as replication weights, is applied for a range of weighted statistics.

Software implementations

Kish's design effect is implemented in various statistical software packages:

- R: survey summary from the survey package.[45] It is also implemented in other R packages (e.g., pew methods[46], and samplesize4surveys[47]).
- Python: design effect from the balance package.[48]
- SAS: Using Proc Survey means.[49]
- Stata: Using the estat post-estimation command after the svy: mean command.[50]
- sudaan.[51]
- WESVAR: calculates Kish's design effect with replacement (SRSWR), i.e. $Deft$.[52]

Notes

a. I.e., that the design effect is the ratio of variances of two estimators, one from a sample with some design and the other from a simple random sample

b. A general formula for the (theoretical) design effect of estimating a total (not the mean), for some design, is given in Cochran 1977.[3]:54

c. The original intention of Kish for $Deft$ was to have it "express the effects of sample design beyond the elemental variability $\frac{S_n^2}{n}$, removing both the unit of measurement and sample size as nuisance parameters". The hope was to have the design effect generalizable (relevant for) many statistics and variables within the same survey (and even between surveys).[5]:55 However, followup works have shown that the design effect depends on the specific sampling design, the outcome, and the statistic of interest (E.g. population total versus the mean). Especially, the $Deft$ depends on the association between some specific outcome with a specific design (e.g. the correlation between y_i and the selection probability p_i).[4]:5 Hence, current literature does not support the generalizability of the $Deft$ across many statistics and outcome measures.

d. As a simple illustration of this, imagine we have clusters of different sizes, and we sample only one cluster (using SRS) and measure all the elements in it. This will lead to EPSEM, but the number of observations we'll get will depend on the cluster size.

e. To be more precise: suppose that S_i is the measure of size for cluster i . One common method of PPS (probability proportional to size) sampling is to sample each cluster with selection probability that is proportional to its size as follows: $P(\text{Selecting cluster } i) = \frac{mS_i}{\sum_{U_i \in U} S_i}$ where m is the number of clusters that we want to sample and U is the frame used for sampling clusters. If we subsampled an equal number, $\bar{n}$, of elements within each sample cluster using some equal probability method, and S_i is the correct number of elements in cluster i, then the selection probability of element j (in some cluster i) will be the same for every element across all clusters (i.e., EPSEM): $\pi_j = \frac{mS_i}{\sum_{U_i \in U} S_i} \frac{\bar{n}}{S_i} = \frac{m\bar{n}}{\sum_{U_i \in U} S_i}$. If S_i turns out not to be the correct size, sampling at the rate of $\frac{\bar{n}}{S_i}$ will still yield EPSEM (equal probability selection method). Notice that if we enumerate (take measurement of) all units in a sample cluster (instead of some fixed number $\bar{n}$, or a fixed proportion $\frac{\bar{n}}{S_i}$), then each unit in cluster i has

the selection probability of the cluster, which will lead to unequal probability of selections between elements of different clusters (i.e., $\pi_j(i) = \frac{mS_i}{\sum_{U_i \in U} S_i}$).

f. For example, say that we assume for each cluster i that its size is S_i, we can sample m clusters with the following probability of selection: $P(\text{Selecting cluster } i) = \frac{mS_i}{\sum_{U_i \in U} S_i}$. And then, we take a fixed number of $\bar{n}$ elements from each cluster. In such a case, if we say that the real cluster size is, say, S_i^*, then the selection probability for each element j taken from cluster i, will be: $\pi_j(j) = \frac{mS_i}{\sum_{U_i \in U} S_i} \frac{\bar{n}}{S_i^*}$. Note that this could be mitigated at the sampling stage if we sample from each cluster using the rate $\frac{\bar{n}}{S_i}$, then the selection probability will be EPSEM (even though the real cluster size was S_i^* and not S_i).

g. This formula would apply only if an equal probability sample were selected in stratum *h* and each element has the same probability of responding.

h. Notice that there is another term called relative variance, which is different. It is the ratio of variance to the mean, while Kish's *relvariance* is the ratio of the variance to the squared mean.

i. In the literature, the sample and the population sizes are sometimes marked as *n* and *N*, and sometimes *m* and *M*. In this article we used *n* and *N*.

j. The formula for Kish's design effect using the original notation:[36]:16

$$Deff_{Kish} = deft^2 = deftu_s^2(1+L) = \left(1+\rho(\bar{b}-1)\right)\frac{n\sum k_j^2}{\left(\sum k_j\right)^2}$$

References

1. Kish, Leslie (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc. ISBN 0-471-10949-5.
2. Heo, Moonseong; Kim, Yongman; Xue, Xiaonan; Kim, Mimi Y. (2010). "Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial". *Statistics in Medicine* **29** (3): 382–390. doi:10.1002/sim.3806. ISSN 1097-0258. PMID 20014353. Archived from the original on 5 January 2013.
3. Sarndal, Carl-Erik; Swensson, Bengt; Wretman, Jan (1992). *Model Assisted Survey Sampling*. Springer. doi:10.1007/978-1-4612-4378-6. ISBN 9780387975283.
4. Park, Inho; Lee, Hyunshik (2004). "Design effects for the weighted mean and total estimators under complex survey sampling". *Survey Methodology* **30** (2): 183–193. ISSN 1492-0921.
5. Kish, Leslie (1995). "Methods for design effects". *Journal of Official Statistics* **11** (1): 55. ISSN 0282-423X.
6. Cochran, William G. (June 1951). "General Principles in the Selection of a Sample". *American Journal of Public Health and the Nations Health* **41** (6): 647–653. doi:10.2105/AJPH.41.6.647. ISSN 0090-0036.
7. Everitt, B.S. (2002). *The Cambridge Dictionary of Statistics* (2nd ed.). Cambridge University Press. ISBN 0-521-81099-X.
8. *Kalton, Graham; Brick, J. Michael; Lê, Thanh (2005). Estimating components of design effects for use in sample design (PDF). Household Sample Surveys in Developing and Transition Countries (Report). New York: Department of Economic and Social Affairs, Statistics Division, United Nations. pp. 95–121. ISBN 92-1-161481-3. ST/ESA/STAT/SER.F/96.*
9. Kish, Leslie (1992). "Weighting for unequal *Pi*". *Journal of Official Statistics* **8** (2): 183–200. ISSN 0282-423X.
10. *Leinster, Tom (18 December 2014). "Effective Sample Size". The n-Category Café.*
11. Wolter, Kirk M. (2007). *Introduction to Variance Estimation* (2nd ed.). Springer. doi:10.1007/978-0-387-35099-8. ISBN 978-0387329178.
12. Frerichs, R. R. (2004). "Equal Probability of Selection". *Rapid Surveys*. unpublished.

13. Valliant, Richard; Dever, Jill A.; Kreuter, Frauke (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer. doi:10.1007/978-1-4614-6449-5. ISBN 978-1-4899-9381-6.

14. Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Nashville, TN: John Wiley & Sons. ISBN 978-0-471-16240-7.

15. Neyman, Jerzy (1934). "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection". *Journal of the Royal Statistical Society* **97** (4): 558–625. doi:10.2307/2342192. ISSN 0952-8385.

16. Dever, Jill A.; Valliant, Richard (2010). "A comparison of variance estimators for post-stratification to estimated control totals". *Survey Methodology* **36** (1): 45-56. ISSN 1492-0921.

17. Kott, Phillip S. (2006). "Using calibration weighting to adjust for nonresponse and coverage errors". *Survey Methodology* **32** (2): 133. ISSN 1492-0921.

18. Holt, D.; Smith, T. M. F. (1979). "Post Stratification". *Journal of the Royal Statistical Society. Series A (General)* **142** (1): 33–46. doi:10.2307/2344652. ISSN 0035-9238.

19. Ghosh, Dhiren; Vogt, Andrew (2002). "Sampling methods related to Bernoulli and Poisson Sampling". *Proceedings of the Section on Survey Research Methods* **2002**: 3569–3570. ISSN 0733-5830.

20. Deville, Jean-Claude; Särndal, Carl-Erik (1992). "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association* **87** (418): 376–382. doi:10.1080/01621459.1992.10475217. ISSN 0162-1459.

21. Brick, J. Michael; Montaquila, Jill; Roth, Shelley (2003). "Identifying problems with raking estimators". *Proceedings of the Section on Survey Research Methods* **2003**: 710–717. ISSN 0733-5830.

22. Keiding, Niels; Clayton, David (2014). "Standardization and Control for Confounding in Observational Studies: A Historical Perspective". *Statistical Science* **29** (4): 529–558. doi:10.1214/13-STS453. ISSN 0883-4237.

23. *Lumley, Thomas (2021-05-25). "How to estimate the (approximate) variance of the weighted mean?". Stack Exchange.*

24. *"What types of weights do SAS, Stata and SPSS support?". UCLA Statistical Consulting Group. 2021. Archived from the original on September 2, 2023. Retrieved September 2, 2023.*

25. Kalton, Graham (1968). "Standardization: A Technique to Control for Extraneous Variables". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **17** (2): 118–136. doi:10.2307/2985676. ISSN 0035-9254.

26. Henry, Kimberly A.; Valliant, Richard (2015). "A design effect measure for calibration weighting in single-stage samples". *Survey Methodology* **41** (2): 315–331. ISSN 1492-0921.

27. *Bock, Tim (24 March 2017). "Design Effects and Effective Sample Size". Displayr.*

28. Gabler, Siegfried; Häder, Sabine; Lahiri, Partha (1999). "A model based justification of Kish's formula for design effects for weighting and clustering". *Survey Methodology* **25**: 105–106. ISSN 1492-0921.

29. Little, Roderick J.; Vartivarian, Sonya (2005). "Does weighting for nonresponse increase the variance of survey means?". *Survey Methodology* **31** (2): 161. ISSN 1492-0921.

30. Liu, Jun; Iannacchione, Vince; Byron, Margie (2002). "Decomposing design effects for stratified sampling". *Proceedings of the Section on Survey Research Methods* **2002**: 2124–2126. ISSN 0733-5830.

31. Spencer, Bruce D. (2000). "An approximate design effect for unequal weighting when measurements may correlate with selection probabilities". *Survey Methodology* **26**: 137–138. ISSN 1492-0921.

32. Park, Inho; Lee, Hyunshik (2001). "The design effect: do we know all about it". *Proceedings of the Section on Survey Research Methods* **2001**. ISSN 0733-5830.

33. Rowe, Alexander K.; Lama, Marcel; Onikpo, Faustin; Deming, Michael S. (2002). "Design effects and intraclass correlation coefficients from a health facility cluster survey in Benin". *International Journal for Quality in Health Care* **14** (6): 521–523. doi:10.1093/intqhc/14.6.521. ISSN 1353-4505. PMID 12515339.

34. *Bland, Michael (2005). "Cluster randomised trials in the medical literature". University of York.*

35. *Ahmed, Saifuddin (2009). "Methods in Sample Surveys" (PDF). Johns Hopkins University Bloomberg School of Public Health. pp. 5–6. Archived from the original (PDF) on 2013-09-28.*

36. *Kish, Leslie (1987). "Questions/Answers" (PDF). The Survey Statistician. Vol. 17. pp. 13–17. ISSN 0214-3240.*

37. Lynn, Peter; Gabler, Siegfried (2005). "Approximations to *b** in the prediction of design effects due to clustering". *Survey Methodology* **31** (1): 101–104. ISSN 1492-0921.

38. Gabler, Siegfried; Hader, Sabine; Lynn, Peter (2005). "Design effects for multiple design samples". *Survey Methodology* **32** (1): 115–120. ISSN 1492-0921.

39. Liu, Jun; Aragon, Elvessa (2000). "Subsampling strategies in longitudinal surveys". *Proceedings of the Section on Survey Research Methods* **2000**: 307–312. ISSN 0733-5830.

40. Park, Inho; Winglee, Marianne; Clark, Jay; Rust, Keith; Sedlak, Andrea; Morganstein, David (2003). "Design effects and survey planning". *Proceedings of the Section on Survey Research Methods* **2003**: 3179–3186. ISSN 0733-5830.

41. Chen, Sixia; Rust, Keith (2017). "An extension of Kish's formula for design effects to two-and three-stage designs with stratification". *Journal of Survey Statistics and Methodology* **5** (2): 111–130. doi:10.1093/jssam/smw036. ISSN 2325-0984.

42. Lohr, Sharon L. (2014). "Design Effects for a Regression Slope in a Cluster Sample". *Journal of Survey Statistics and Methodology* **2** (2): 97–125. doi:10.1093/jssam/smu003. ISSN 2325-0984.

43. Zins, Stefan; Burgard, Jan Pablo (2020). "Considering interviewer and design effects when planning sample sizes". *Survey Methodology* **46** (1): 93–119. ISSN 1492-0921.

44. Potter, Frank; Zheng, Yuhong (2015). "Methods and issues in trimming extreme weights in sample surveys". *Proceedings of the Section on Survey Research Methods* **2015**: 2707–2719. ISSN 0733-5830.

45. Lumley, Thomas (2004). "Analysis of Complex Survey Samples". *Journal of Statistical Software* **9** (1): 1–19. doi:10.18637/jss.v009.i08. ISSN 1548-7660.

46. *Pew Research Center. "pewmethods". GitHub. Retrieved November 28, 2023.*

47. *Gutierrez Rojas, Hugo Andres (January 17, 2020). "samplesize4surveys". The Comprehensive R Archive Network (CRAN). Retrieved November 28, 2023.*

48. *Sarig, Tal; Galili, Tal; Eilat, Roee (2023). "balance -- a Python package for balancing biased data samples". arXiv:2307.06024 [stat.CO].*

49. *Buskirk, Trent D. (2011). Estimating Design Effects for Means, Proportions and Totals from Complex Sample Survey Data Using SAS® Proc Surveymeans (PDF). Midwest SAS Users Group Conference 2011. Saint Louis, MO: Saint Louis University School of Public Health. pp. 1–13. Archived from the original (PDF) on May 11, 2015. Retrieved November 28, 2023.*

50. *"Survey Data Analysis in Stata 17". UCLA Statistical Consulting Group. 2021. Archived from the original on June 7, 2023. Retrieved November 28, 2023.*

51. *"DESCRIPT Example 1" (PDF). RTI International. Retrieved November 28, 2023.*

52. *Choudhry, G. Hussain; Valliant, Richard (2002). WesVar: Software for complex survey data analysis (PDF). Statistics Canada Symposium. Ottawa: Statistics Canada. Retrieved November 28, 2023.*