Environmental Health Perspectives
Vol 90, pp. 215-222, 1991

Assessing, Accommodating, and Interpreting the Influences of Heterogeneity

by Thomas A. Louis*

Heterogeneity, ranging from measurement error to variation among individuals or regions, influences all levels of data collected for risk assessment. In its role as a nemesis, heterogeneity can reduce the precision of estimates, change the shape of a population model, or reduce the generalizability of study results. In many contexts, however, heterogeneity is the primary object of inference. Indeed, some degree of heterogeneity in excess of a baseline amount associated with a statistical model is necessary in order to identify important determinants of response. This report outlines the causes and influences of heterogeneity, develops statistical methods used to estimate and account for it, discusses interpretations of heterogeneity, and shows how it should influence study design. Examples from dose-response modeling, identification of sensitive individuals, assessment of small area variations and meta analysis provide applied contexts.

Introduction

Without heterogeneity, or variation, there would be no potential for gaining scientific information. Yet, unappreciated heterogeneity can degrade or distort scientific interpretations.Therefore, though heterogeneity may play a negative role in some situations, it is our lifeblood in others. This report discusses issues, techniques, and examples related to assessing, accommodating, interpreting, and controlling the effects of heterogeneity. Though no abstract definition of heterogeneity is satisfactory for all settings, we shall propose operational definitions that are relevant to specific scientific and practical questions. We do not go into great detail on any specific issue or example, rather we provide an overview and key into a wide literature.

Heterogeneity can be defined in many ways with the most inclusive being variation in general. More restrictive definitions help structure our approach. Heterogeneity as variation in excess of a baseline model provides an important framework. For example, Knuiman et al. (*1*), Margolin et al. (*2,3*), and McCann et al. (*4*) show that intra- and interlaboratory variation in the number of revertants in the Ames test far exceeds that predicted from the Poisson distribution, and we have a considerable amount of unexplained variation. Interpretations of the Ames test should be based on the true variability, so identifying the excess has direct value. Equally important is an attempt to explain excess variation. Much of this may be inexplicable, but some may be accounted for by covariates such as variation in the growth medium and incubation temperature. Explaining some of the excess allows improved experimental design, and, with a validated model that relates covariates to outcomes, gives the ability adjust results and to put outcomes on a common basis. This adjustment can reduce variability and bias.

The problem of defining and identifying sensitive individuals provides another instructive example. As Bailar and Louis (*5*) discuss for lung responsiveness to challenge, smokers as a group may be considered sensitive relative to nonmokers, yet some smokers (even in the same smoking rate category) are more sensitive than others. This example identifies the question of scoping the assessment of heterogeneity. One needs to group study units that can be considered homogeneous (follow a baseline model). This decision is made using a combination of scientific and practical considerations.

A third class of examples, called measurement error models, shows the importance of identifying and accounting for heterogeneity. Consider the effect of diastolic blood pressure on risk of stroke. Blood pressure measurements are made with error (*6,7*), and it is well known that estimated regression slopes are attenuated relative to the relation between true blood pressure and risk of stroke. Of course, the reported regression slope is correct for the question of how stroke risk relates to measured blood pressure. But this slope does underestimate the true influence of blood pressure on stroke. MacMahon et al. (*8*) show, for example, that the reported slope should be increased by approximately 60% to estimate the underlying relation. This increase has profound policy and health implications.

Identifying this attenuation of slope is important when combining evidence over a variety of studies through a meta analysis or overview (*9*). If each study uses a different measurement system (for example, taking the average of a different number of blood pressure readings), reported slopes will need adjustment before combining. Failure to do so can result in observed heterogeneity of slopes in excess of that predicted by sampling variability (the baseline model). In this case, the unexplained

*Division of Biostatistics, University of Minnesota, School of Public Health, Mayo Box 197, Minneapolis, MN 55455.

variation can be explained and accommodated by an error-in-variables model.

One more class of examples serve to introduce another feature of unappreciated heterogeneity. Vaupel and Yashin (*10*) and Yashin (*11*) discuss "heterogeneity's ruses," where the shape of the hazard for death in a population is different from that for any individual. For example, each individual may have a constant hazard (exponential distribution), but the population curve will show a decreasing hazard. Relatively speaking, the frail die out early, leaving the hardy (low hazard) to live. Again, at the population level, the decreasing hazard is appropriate, but if one wishes to study aging or policy impacts, an understanding of the ruses and an attempt using data and theory to uncover them is vital. These ruses are just another example of Simpson's paradox.

The policy issue in survival models is easily seen in studies of the effects of smoking cessation on the risk of lung cancer. A multistage model predicts that the excess risk continues to increase even after an individual stops smoking (though the increase is less than for an individual who continues to smoke), whereas many data sets show that the excess decreases. The decrease may be true, and surely is for the population, but each individual's excess risk may indeed increase, but heterogeneously.

The facts discussed so far, suggest several definitions of heterogeneity, including general variability, variability in excess of a baseline model, variance components, measurement error, variation in latent parameters. To these definitions we can add heterogeneity induced by variation in data analytic approaches. In the end, heterogeneity is a vague concept and is best appreciated through a series of examples.

Two-Stage Models

Basic

The two-stage model provides a convenient way to represent variation in excess of a baseline model. In this model we have stage II: a parameter (vector) is sampled from a distribution; and stage I: data are generated from a sampling distribution, conditional on the parameter. This two-stage sampling process can be repeated for each experimenal unit (e.g., clinics, petri dishes, small geographic areas), and underlies Bayes and empirical Bayes analysis. For concreteness, consider a stage II governed by a prior distribution (G) that is Gaussian and that stage I is Gaussian with a known variance (*12,13*). Let k denote experimental unit, $k = 1, \ldots, K$, and the basic model becomes:

$$\theta_1, \theta_2, \ldots, \theta_K \text{ are iid } N(\mu, \tau^2)$$
$$Y_k \mid \theta_k \sim N(\theta_k, \sigma^2).$$

This model can be extended in a wide variety of ways, including having repeated sampling of Y values for each θ_k, allowing distributions different from the Gaussian, using unequal sampling variances, and introducing a regression structure for the Y so that the θ values come from different, but related distributions. Continuing with this basic model we have:

$$Y_1, \ldots, Y_K \text{ are (iid) } N(\mu, \sigma^2 + \tau^2)$$

and the Y_k values are overdispersed relative to the sampling distribution with variance σ^2, but it may be explicable.

This model is the model II or random effects ANOVA, and tests of the null hypothesis are asking if $\tau^2 = 0$. If $\tau^2 > 0$, then we have variation not accounted for by the baseline model and can look for explanations through covariance adjustment. If useful covariates are present, excess (unexplained) variation will be reduced by the adjustment. Notice that the notion of excess variation depends on the stage I model (the sampling distribution). If we find in a data set that the sample variance:

$$\frac{1}{K-1} \Sigma (Y_k - \overline{Y})^2 = 10,$$

and we know (assume) that $\sigma^2 = 6$, then a natural estimate for τ^2 is 4. If, however, we assume $\sigma^2 = 9$, then $\tau^2 = 1$. In practice, we need either direct information on sampling variation (through replication within units) or reliable assumptions (e. g., use of arcsine transformed binomial data) to identify τ^2.

For another example of a two-stage model, consider the case where the sampling distribution is Poisson (*14,15*). Then,

$$Y_k \mid \theta_k \sim \text{Poi}(\theta_k), \; E(Y_k \mid \theta_k) = \theta_k, \; V(Y_k \mid \theta_k) = \theta_k.$$

Frequently, variability among the Y values is greater than that predicted by their sample mean, and a two-stage model is suggested (*1–3*). For example, assuming that the θ values are gamma with mean, μ, and squared coefficient of variation α^{-1}, we find that the Y_k values are negative binomial with:

$$E(Y_k) = \mu, \; V(Y_k) = \mu(1 + \frac{\mu}{\alpha}),$$

allowing for overdispersion. Even if we do not accept the two-stage hierarchy, the negative binomial allows more flexible modeling than does the Poisson.

Consequences

Going back to the Gaussian example, let us consider estimation of the mean (μ), and estimation of the individual θ values. For the former, allow n replications for each θ. The optimal estimate of μ is $\overline{Y}$, the mean of all observations. The variance of $\overline{Y}$ consists of two components: that induced by the prior and that induced by the sampling distribution. We have:

$$\overline{Y} = \frac{1}{K} \Sigma \overline{Y}_k, \; V(\overline{Y}) = \frac{\sigma^2}{Kn} + \frac{\tau^2}{K}.$$

Notice that increasing either K or n will reduce the first term, but the second term is controlled by K, the number of θ values that we sampled (primary units). In fact, if $n = \infty$, so that we have perfect knowledge of each θ_k, we still have uncertainty in estimating μ.

A fixed effects analysis, where we wish to estimate $\overline{\theta}$, the mean of the θ values generating these data, assumes no connection among them. The random effects analysis strives to make an inference about μ, which is broader than to the current experimental

units, and this broader inference comes with increased variability. To get an idea of the increase, consider the ratio of the variability for the random effect (RE) versus fixed effect (FE) analysis. We have:

$$\frac{V_{RE}}{V_{FE}} = 1 + (n - 1)\,\rho,$$

where $\rho = \tau^2/(\sigma^2 + \tau^2)$ is the intraclass correlation. If n = 1 or $\rho = 0$ ($\tau = 0$), there is no variance inflation, otherwise it can be considerable. This ratio is called the design effect. In the extreme case with $\rho = 1$, n observations on an experimental unit are no more informative (variance reducing) than one observation and V_{RE} is n times V_{FE}.

The design effect and generalizations thereof can be used to determine how many repeat measurements on a unit (n) and how many units (K) are needed to produce a desired accuracy. Setting n = 1 minimizes the number of observations, but generally repeat observations on a unit are more relevant or less expensive than adding units, so an n > 1 usually produces the optimal solution under resource constraints.

The decision to use fixed or random effects depends on inferential goals, and differences in goals underlie the current controversy over when to use meta analysis (*16*). The random effects approach produces considerably larger standard errors on an estimated treatment effect than does the fixed effects analyses. Peto and colleagues argue in favor of inferences limited to the meta-analyzed studies and used fixed effects (*17*). Others desire a broadened inference to the population of similar studies, and this promotes random effects (*18*). The narrow inference has a well-defined reference population but does not easily generalize. The broad inference is possibly more relevant, but the target population is somewhat vague.

Irrespective of these scoping issues, the stage II variance component can be of independent interest. Consider the metaanalysis of the effect of coaching on scholastic aptitude test (SAT) scores conducted by Laird and DerSimonian (*19*). Table 1 shows that estimated effects are greater for uncontrolled than for controlled studies (likely because much of the coaching effect is really regression to the mean), reports the standard error of the estimate (the SE depends both on the number of studies and the sample size for each study), and shows that the stage II variability is greater for uncontrolled studies. We can see that unexplained variation is greater for the uncontrolled studies (it is likely that they are performed under a wide variety of conditions) and under the Gaussian assumption can get an idea of the variation in true coaching effect. True coaching effect for uncontrolled studies can be expected to vary according to a $N(41, 25^2)$ distribution, while for matched/randomized studies it follows something close to a $N(10, 3^2)$. This information is of policy interest. For example, it is virtually impossible for there to be a negative coaching effect in the randomized approach applied to populations similar to those used in the current studies.

Table 1. The effect of coaching on SAT scores (*19*).

	Coaching effect		
Parameter	Uncontrolled	Controlled	Matched/ randomized
$\hat{\mu}$	41	15	10
SE ($\hat{\mu}$)	10	5	4
$\hat{\tau}$	25	14	3

Table 2. Meta analysis of vinyl chloride from Beaumont and Breslow (*20*).[a]

Site	Relative risk	*p*-Values H_0:RR = 1	Heterogeneity
Liver	5.2	0.0001	0.002
Brain	1.7	0.01	0.100
Lung	1.1	Not significant	0.060

[a]There were nine studies.

Tests for heterogeneity in a Gaussian framework ask if $\tau^2 = 0$. In general, these tests are less informative than reporting an estimate, where one can determine if the excess variation is a threat to interpretation or generalization. Consider, for example, the meta analysis conducted by Beaumont and Breslow (*20*) on the cancer risk from vinyl chloride. Table 2 shows their summary, indicating statistically significant relative risks for liver and brain, but not for lung tumors. For liver, the test for heterogeneity conclusively shows that the studies are not estimating a common relative risk (there is heterogeneity). An estimated variance component would allow one to see if variation in relative risks is sufficient to produce values below 1 (a qualitative interaction) or simply variation all to the right of 1 (a quantitative interaction). The former causes problems in interpretation, the latter does not. Both set the stage for finding explanations for the excess variation.

Example

An early two-stage analysis of water contamination was published by Von Mises (*21*). In each of 3420 sampling sites, 5 samples were taken and the number of contaminated samples recorded. Table 3 gives the data and expected frequencies (rounded) under the binomial distribution, computed with the estimated contamination probability of 0.025. For this event probability, the binomial assumption predicts far too few occurences of 2 to 5, and a two-stage variance components model can capture the excess variation among geopraphic areas. The sample variance of the observed contaminations is 0.1885, which is greater than the 0.1219 predicted from the binomial [0.1219 = 5(0.025)(0.975)], suggesting overdispersion.

The beta distribution is a common model for stage II, and we parameterize it by the mean, μ, and intraclass correlation $(M + 1)^{-1}$. Specifically, with θ the binomial parameter:

$$E(\theta) = \mu, \quad V(\theta) = \mu(1 - \mu)/(M + 1).$$

From the Von Mises data, we obtain $\hat{\mu} = 0.025$, M = 6.2. The 6.2 shows high *a priori* variation, whereas an $M \to \infty$ would indicate no prior variation.

We can go further with this example and free ourselves from assuming a specific parametric shape for the prior by using a nonparametric estimate. Laird (*22*) introduced the algorithm, and we obtain a discrete prior with masses (0.606, 0.261, 0.092, 0.022, 0.018) at mass points (0.012, 0.021, 0.029, 0.052, 0.408). This nonparametric approach allows for flexible modeling, and when its properties are better developed, it should become a standard approach to variance component problems.

Table 3. Distribution of contaminated samples in 3420 sampling sites (21).

Number contaminated	Frequency	Expected
0	3086	3017
1	279	383
2	32	19
3	15	0.5
4	5	0
5	3	0

Transforms

Sometimes heterogeneity can be reduced or eliminated by a data transform. Consider data from a multicenter clinical trial analyzed by the model:

$$Y_{jk} = a_k + b_k T_j + e_{jk},$$

where $T_j = 0$ for $j = 1$ (treatment 1), $T_j = 1$ for $j = 2$ (treatment 2), k denotes the clinic, the a values and b values are random effects, and the Y values are treatment means. If the data are actually log-normal, the proper model is:

$$\log(Y_{jk}) = a_k^* + b_k^* T_j + e_{jk}^*$$

with $b_k^* \equiv b$, we have the situation where $V(b_k) < 0$ but $V(b_k^*) = 0$. This treatment-by-clinic random interaction effect in the untransformed scale is induced by clinics having different baseline responses (the a_k^* vary) and the model being misspecified. Random effects models in the untransformed scale can pick up the interaction induced by misspecification and serve to make the standard analysis more robust. Notice that interaction induced by these transformation models is never qualitative (all the b_k are of the same sign) so combining evidence is scientifically valid. The random-effects model delivers an estimated variance for the treatment effect that accounts for the heterogeneity.

Estimating Individual Components

Now, let us consider estimating the underlying mean for individual units (the θ_k for the Gaussian example). If we knew μ and τ^2, the distribution of θ_k given Y_k (the posterior distribution) is Gaussian with:

$$E(\theta_k | Y_k) = \mu + (1 - B)(Y_k - \mu)$$
$$V(\theta_k | Y_k) = (1 - B)\sigma^2,$$

where $B = \sigma^2/(\sigma^2 + \tau^2)$. Notice that the observed value (Y_k) is shrunken towards the prior mean (μ) by an amount that depends on the relative sizes of the prior and sampling variance and that the variance is less than the sampling variance σ^2. If σ^2 is small (e.g., if Y_k is an average of several replicates), then B is small and very little shrinkage takes place. Similarly, if τ^2 is relatively large (there is very little information *a priori*) we have little shrinkage. Indeed, if $B = 0$, the conditional expectation of θ_k is Y_k, the usual result. If σ^2 is relatively large, considerable shrinkage occurs, stabilizing the estimate at the expense of adding some bias. These relations express the general guideline of modeling where for small sample sizes control of variability dominates analytic goals. For large sample sizes bias reduction dominates. Striking the appropriate tradeoff comes from scientific theories, formal models, and cross-validation methods (*23*).

When we do not know μ and τ^2, they can be estimated from the data (*13*), producing empirical Bayes estimates. Basic estimates used are:

$$\hat{\mu} = \overline{Y}$$
$$\hat{\tau}^2 = \left\{\frac{1}{K-1}\Sigma(Y_k - \overline{Y})^2 - \sigma^2\right\}^+,$$

where + sets negative numbers to 0. Using these empirical Bayes estimates, where information from all units is used to estimate a single unit's mean, can improve estimation performance. Applications by Efron and Morris to toxoplasmosis incidence (*24*), Laird and Louis to carcinogenicity testing (*23*), Rubin to law school admissions criteria (*26*), Louis (*27*) and Tukey (*28*) to histogram estimates, Clayton and Kaldor (*29*) to relative risks, Stroud (*30*) to small area analysis, Hui and Berger to longitudinal studies (*31*), DerSimonian and Laird to clinical trials (*32*), and Wiley et al. (*33*) to AIDS transmission, a long history in actuarial science and examples in this article indicate the rich variety of applications. The success of the approach is due to accounting for variation in underlying parameters while shrinking extreme results towards a group mean, essentially accounting for regression to the mean.

This improvement in estimates carries over to improved confidence intervals in that they attain the nominal coverage but are of shorter length than the classical intervals (*13,34*). This empirical Bayes advantage holds even when the intervals are broadened to account for uncertainty in estimating the prior distribution. With the development of calibrated confidence intervals, the empirical Bayes approach produces inferences ideally suited to risk assessment investigations when data from related sources are available.

Consequences of Heterogeneity

Bias

We have seen that accounting for heterogeneity can adjust standard errors, provide a valid basis for generalization, and improve estimation of individual parameters. Ignoring the heterogeneity can have dire consequences. Cox (*15*) considered the Gamma-Poisson example. If we wish to estimate the mean event rate from several units, the mean $\overline{Y}$ is totally efficient, but we do need to account for heterogeneity to get a proper standard error. However, consider estimating a nonlinear function of the mean μ; for example e^μ. The estimate $e^{\overline{Y}}$ will not have expectation e^μ even for large sample sizes; it is inconsistent. A consistent estimate requires accounting for the heterogeneity.

Correlation

Discovery and accounts of heterogeneity are key ingredients for the analysis of longitudinal data (*35,36*). At the most basic level, consider the correlation between lung function measurements (the forced expiratory volume) in adults at 3-year intervals. With no covariates the correlation is about 0.90, but adjusting for age, height, and gender brings the correlation down

to about 0.80. Some of what used to be unexplained (co)variation has been explained. It is important to note that this covariation is measured using residuals from the model producing expectations, not from the raw data.

Unexplained (co)variation produces a variety of phenomena in longitudinal data. Consider relating adjacent residuals by plotting e_{t+1} versus e_t, where the e values are residuals from a model. A plot with slope 1 represents tracking where an individual's deviation from the population prediction tends to stay put. A slope less than 1 indicates regression to the mean, where the subsequent deviation tends to be less extreme than the current value. A slope greater than 1 indicates the horse race, where residuals tend to increase in absolute value. The name derives from the idea that horses in the lead tend to be running fastest and increase their lead. Of course, a plot showing random scatter with slope 0 indicates no association.

As one includes additional, effective covariates in the model, these residual plots change, sometimes with changed slope, usually with less extreme relations. Again, as the model explains additional (co)variation, we change the structure. For example, one explanation for the horse race comes from assuming that covariates not yet in the model (e.g., smoking), influence the rate of change in lung function. Without the covariate in the model, smokers have the fastest decline and the lowest lung function. Their lead increases over time. Once the smoking covariate is included, residuals compare smokers to smokers, and residuals have a chance to be both positive and negative. Though there still may be a horse race phenomenon, it will be less and could be reduced further by including additional predictors of decline. When to stop adding predictors depends on the available information, subject area knowledge, and statistical art and science.

Errors in Variables

Unappreciated heterogeneity can influence scientific and policy conclusions by disguising true parameter values and relational shapes. Berkson provided the classic example, where we are relating variables through a linear model. For specificity, consider the model: toxicity = a + b • dose, where we take measurements on dose and toxicity. Consider a large sample, so that statistical variation in parameter estimates is not the issue. If dose is measured without error, then b relates true dose to toxicity but if measured dose is a random deviation from true dose, then b will be closer to zero (attenuated) relative to the slope appropriate for true dose. Yet, it is the appropriate value for relating observed dose to toxicity.

Since all we ever deal with are observed values, why should we care about this apparent bias? We care for at least two reasons: (*a*) If different experiments are performed with different measurement accuracies, then failure to account for these differences will produce apparently different slopes. Any meta analysis of experimental results will require adjustment to a common basis. (*b*) Possibly more importantly, even though the unadjusted slope is appropriate for relating measured dose to response, it is inappropriate when making policy recommendations. For example, many studies report on the positive relation between blood pressure and heart attack risk. Since blood pressure generally is measured with considerable variation, the effectiveness of blood pressure control is underestimated. Since policies are aimed at reducing true blood pressure, the adjusted slope is relevant.

Many authors have discussed the consequences of errors in variables and methods for reducing the effects through design and analyses (*6,7,37–40*). Design considerations include averaging repeated measures to reduce heterogeneity. Analyses that explicitly or implicitly de-attenuate regression slopes are effective in performing the necessary adjustments. In studies relating a risk factor to a response in the presence of a confounder, Kupper (*6*) shows that unreliability in measurement of the confounder can be more damaging than unreliability in the risk factor, sometimes producing a sign change between the estimated slope and the slope appropriate for true confounder-adjusted risk. Extreme care is needed in defining research questions, designing, and analyzing studies.

The effects of unaccommodated heterogeneity can be more dramatic than slope attenuation. Consider the linear model

$$Y = a + bx + \text{error}$$

where x is the true regressor. Let the observed regressor (X) conditional on the true regressor be distributed as a log-normal variable with mean x. Then, the regression using the observed regressor is:

$$y = a + bX^p + \text{error} ,$$

where $p < 1$, so a linear relation is converted to nonlinear. Bailar et al. (*41*) suggest that this type of phenomenon (operating as variation in true slope from rodent to rodent) may produce the apparent lack of conservatism for linear extrapolation of dose-response relations. The dose-response curve for vinyl chloride is a classic example.

When we start with a nonlinear model, such as a logistic regression or a multistage model for carcinogenicity, heterogeneity due to noisy regressors or due to variations in true slopes from unit to unit (the compound model) change the shape of the relation. Techniques are available for estimating and adjusting (*39,41*) but more development of numerical methods is required.

Survival Analysis

Survival analysis gives a particularly transparent view of the influence of heterogeneity (*10,11,43–47*). Recall that the hazard, or force of mortality, is defined as:

$$\mu(t)dt = \text{pr}(\text{death in } (t,t+dt) \mid \text{alive at } t).$$

With S(t) the survival curve:

$$\mu(t) = -\frac{d}{dt}\log(S(t))$$

or

$$S(t) = \exp[-U(t)],\ U(t) = \int_0^t \mu(t)dt .$$

Assume that each individual in a population has a hazard depending on a parameter θ that multiplies a baseline hazard (proportional hazards), and that θ varies from person to person according to a probability distribution G (again, the two-stage model). Then, for the population:

$$S_G(t) = \int_0^\infty e^{-\theta H(t)}\, dG(\theta)$$

and

$$\mu_G(t) = h(t) \cdot E_G(\theta \mid T > t)\ ,$$

where h is the basline hazard, and H is its integral. It can be shown the $E_G(\theta \mid T > t)$ is decreasing in t so that the shape of $\mu_G(t)$ is different from that of h(t).

Vaupel and Yashin (*10*) give several examples of this difference in shape. The easiest example has $h(t) \equiv 1$ and G putting mass on two points $\theta_1 < \theta_2$. Then, though each individual has a constant hazard, the population hazard decreases from θ_2 to θ_1 as a function of t. Intuitively, those alive at a large t are not representative of the original population, but overrepresent those with the smaller θ (θ_1). Individuals with θ_2 tend to die off early. As in errors in variables regression, $\mu_G(t)$ is the appropriate population curve, but it may be deceptive in terms of determining appropriate policies or their effects. For example, it may be true that it is good for every individual in a population to stop smoking (all age-specific hazards decrease), but that after some follow-up interval of time, some age-specific hazards are higher than they would have been without the intervention. Some of the frail individuals who would have died earlier if they had continued to smoke have been sent to deaths at later ages. This is another manifestation of Simpson's paradox; aptly put as an intervention that is good for men; good for women; but appears bad for people. Vaupel and Yashin (*10*) call this the apparent failure of success and remind us carefully to consider heterogeneity when designing and evaluating programs.

Risk Assessment

The decision on using historical controls in the analyses of the carcinogen bioassay illuminates several issues related to heterogeneity. Consider a bioassay comparing lifetime tumor rates between an exposed and control group of rodents, each with 50 rodents. Table 4 presents typical data when all tumors occur in the exposed group. The Fisher's exact one-sided p-value is approximately $(0.5)^x$, so that if x is less than 5, the p-value will exceed the usual level for statistics significance. The p-values for X = 0,1,2,3,4,5 are 1.000, 0.500, 0.247, 0.121, 0.059, and 0.028. However, in many situations pathologists will report that even an X = 3 is biologically significant because in a long series of experiments virtually no tumors of the type being considered have been found in the control group.

This dissensus between the statistical procedure and scientific opinion can be explained and rectified by a two-stage model where the control rate for the current experiment is considered to be sampled from a prior distribution. In the case we are considering, this prior puts almost all weight on a control rate of 0 and is equivalent to increasing the control sample size. For example, consider the situation if Table 4 were modified to have 450 control rodents, none with the tumor in question. Then, the one-sided Fisher's exact test is approximately $(0.1)^x$ and three tumors produce $p = 0.001$, a statistically significant result.

Table 4. Hypothetical results from a carcinogen bioassay.[a]

	Control	Exposed	Total
Tumor	0	x	x
No tumor	50	50 − x	100 − x
Total	50	50	100

[a]The one-sided Fisher's exact p-value is approximately $(0.5)^x$.

In general, the use of two-stage models for this experiment is more complicated. Dempster et al. (*48*) Tamura and Young (*49*), and the references thereof explain approaches. Formalized use of the historical data will help resolve controversies such as those that surrounded the assay for DMT (dimethylterephthalate), where the data showed 2%, 16%, and 27% lifetime incidence of alveolar/broncheolar adenomas and carcinomas in male mice for the control, low-, and high-dose groups ($p < 0.0001$), but previous control groups in the same laboratories had rates of 10%, 13%, and 18% (*50*). Refinements are needed to incorporate time-until-tumor and cause-of-death information, and the approach should be included as a formal method of focusing discussion of bioassay results. As evidenced by Freedman (*51*), the carcinogen bioassay generates uncertainty and controversy more broadly.

Hierarchical models incorporating complicated variance component structures have been used to relate data from seemingly unrelated human studies and to formalize interspecies extrapolations of Risk. DuMouchel and Harris (*52*) present an analysis of the health effects of environmental emissions, and Laird (*53*) investigates the thyroid cancer risk of ionizing radiation. Both of these analyses carefully lay out assumptions and study sensitivity of results to changes in assumptions. The formal approach provides explicit documentation of methods and helps focus discussion of modifications. These models will always be augmented by expert opinion and political considerations when used as input to risk assessment and control, but they serve an extremely valuable role, going beyond more descriptive approaches such as those of Crouch and Wilson (*54*).

Scoping

When assessing or accommodating heterogeneity, the analysis frame is extremely important. One needs to decide on the basic unit of analysis (e.g., the individual, the publication, the small geographic area), the baseline model (e.g., logistic dose response, Poisson counts), the form of heterogeneity (e.g., gamma, Gaussian), and the type of units to be admitted to the analysis. We refer to the types of units analyzed as scoping, and it concerns defining the types of units that can be expected to be related by the heterogeneity distribution. Too broad a scoping will produce estimated prior variation so large as to unduly inflate the standard error of parameter estimates and reduce the advantage in estimating unit-specific parameters of combining evidence over units. Too narrow a scoping produces unstable estimates of the prior and constrains generalizations of the findings.

One approach to scoping includes units that are thought to have

parameters sampled from a common prior distribution. A more general approach allows covariates to adjust the prior distribution. For example, in performing a small area analysis of disease incidence (*29*), the prior can be adjusted for age, gender, and risk factors by building a regression model for the prior mean. This approach allows for broadened inferences and is the model-based method for explaining unexplained variation.

Use of historical controls in the bioassay provides a good example of the issues. The analyst needs to decide what controls to include. Should they be from the same laboratory or several laboratories? Should controls from tests performed several years previously be included or only recent controls? Should control rates from experiments read by the same pathologist be included? Each choice influences the effect of including historical controls and the operating characteristic of the test procedure.

Amato and Lagakos (*55*) analyze the effects of disagreement among pathologists when characterizing the tumors in rodents used in a carcinogen bioassay. They show how dose-response curves vary from pathologist to pathologist even when the same slides are read. For bladder and liver tumors, dose-response curves from different pathologists essentially do not cross, implying that some individuals have a substantially higher rate of positives at all doses. This variation in call rate produces variation in the statistical power of the bioassay and limits its generalizability. The variation invites explanation, with likely explanations including variation in individual pathologists' perceived prevalence of tumor types and variation in the degree to which they incorporate the consequences of false positives and false negatives into their decision rule. If the causative factors can be isolated, pathology protocols can be modified to reduce this inter-rate variation.

Summary

We have seen that heterogeneity is the foundation of statistical science and that its identification and accommodation is centrally important to the design, conduct, analysis, and interpretation of statistical studies. Key issues include determining baseline models and defining analytic goals. If the goals are primarily to make inferences at the population level, then heterogeneity modeling is quite robust and generally serves to expand the flexibility of population models to represent expectations and variances. If, however, inferences are directed at the unit-specific level, considerable care is needed in specifying baseline models and forms for heterogeneity. Commonly, true replications at the unit level are unavailable, and models will be based on a combination of scientific reasonableness and statistical/mathematical convenience. Usually, the specific forms chosen are not uniquely best for the observed data, and careful interpretation coupled with sensitivity analysis is required.

Effective assessments, accommodations, and interpretations of heterogeneity require effective team work among statisticians and other scientists tuning the approach to the application. The scientifically challenging and societally important problems in quantitative risk assessment provide fertile ground for developing and applying methods that will increase scientific understanding and improve the public health.

This work was supported by grant DMS 8402720 from the National Science Foundation and a grant from the Sloan Foundation. The author thanks Margaret Andrews for preparing the manuscript.

REFERENCES

1. Knuiman, M. W., Laird, N. M., and Louis, T. A. Inter-laboratory variability in Ames assay results. Mutat. Res. 180: 171–182 (1987).
2. Margolin, B. H., Kaplan, N., and Zeiger, E. Statistical analysis of the Ames Salmonella/microsome test. Proc. Natl. Acad. Sci. USA 78: 3779–3783 (1981).
3. Margolin, B. H., Risko, K. J., Shelby, M. D., and Zeiger, E. Sources of variability in Ames Salmonella typhimurium tester strains: analysis of the International Collaborative Study on "Genetic Drift." Mutat. Res. 130: 11–25 (1984).
4. McCann, J., Horn, L., and Kaldor, J. An evaluation of Salmonella (Ames) test data in the published literature: application of statistical procedures and analysis of mutagenic potency. Mutat. Res. 134: 1–47 (1984).
5. Bailar, J. C., and Louis, T. A. Statistical concepts and issues. In: Variations in Susceptibility to Inhaled Pollutants, Identification, Mechanisms, and Policy Implications (J. D. Brain, B. D. Beck, A. J. Warren, and R. A. Shaikh, Eds.), The Johns Hopkins Press, Baltimore, MD, 1988, pp. 30–55.
6. Kupper, L. L. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. Am. J. Epidemiol. 120: 643–648 (1984).
7. Liu, K. Measurement error and its impact on partial correlation and multiple linear regression analyses. Am. J. Epidemiol. 127: 864–874 (1988).
8. MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbot, R., Godwin, J., Dyer, A., and Stamler, J. Blood pressure, stroke, and coronary heart disease. Part 1: Prospective observational studies corrected for the regression dilution bias. Lancet 335: 765–774 (1990).
9. Louis, T. A., Fineberg, H. V., and Mosteller, F. Findings for public health form meta-analysis. Ann. Rev. Public Health 6: 1–20 (1985).
10. Vaupel, J. W., and Yashin, A. I. Heterogeneity's uses: some surprising effects of selection on population dynamics. Am. Stat. 39: 176–185 (1985).
11. Yashin, A. I., Manton, K. G., and Vaupel, J. W. Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. Theo. Popul. Bio. 27: 154–175 (1985).
12. Casella, G. An introduction to empirical Bayes data analysis. Am. Stat. 39: 83–87 (1985).
13. Morris, C. Parametric empirical Bayes inference: theory and applications. J. Am. Stat. Assoc. 78: 45–59 (1983).
14. Breslow, N. E. Extra-Poisson variation in log-linear models. Appl. Stat. 33: 38–44 (1984).
15. Cox, D. R. Some remarks on overdispersion. Biometrika 70: 269–274 (1983).
16. Colton, T., Freeman, L. S., and Johnson, A. L., Eds. Proceedings of the Workshop on Methodological Issues in Overviews of Randomized Clinical Trials, May 1986. Stat. Med. 6(3) (1987).
17. Peto, R. Why do we need systematic overviews of randomized trials? Stat. Med. 6: 233–240 (1986).
18. Discussion of "Why do we need systematic overviews of clinical trials?" Stat. Med. 6: 241–244 (1986).
19. DerSimonian, R., and Laird, N. M. Evaluating the effectiveness of coaching for SAT exams: a meta-analysis. Harvard Ed. Rev. 53: 1–15 (1983).
20. Beaumont, J. J., and Breslow, N. E. Power considerations in epidemiologic studies of vinyl chloride workers. Am. J. Epidemiol. 114: 725–734 (1981).
21. von Mises, R. On the correct use of Bayes' formula. Ann. Math. Stat. 13: 156–165 (1942).
22. Laird, N. M. Non-parametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc. 73: 805–811 (1978).
23. Efron, B., and Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37: 36–48 (1983).
24. Efron, B., and Morris, C. Stein's paradox in statistics. Sci. Am. 236: 119–127 (1977).
25. Laird, N. M., and Louis, T. A. Empirical Bayes confidence intervals for a series of related experiments. Biometrics 45: 481–495 (1989).
26. Rubin, D. B. Using empirical Bayes techniques in the law school validity studies. J. Am. Statist. Assoc. 75: 801–827 (1980).
27. Louis, T. A. Estimating a population of parameter values using Bayes and empirical Bayes methods. J. Am. Stat. Assoc. 79: 393–398 (1984).
28. Tukey, J. W. Named and faceless values: an initial exploration in memory of Prasant C. Mahalanobis. Sankhya 36: 125–176 (1974).
29. Clayton, D., and Kaldor, J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43: 671–681 (1987).
30. Stroud, T. M. F. Bayes and empirical Bayes approaches to small area estimation. In: Small Area Statistics: An International Symposium (R. Platcek, J.N.K. Rao, C.E. Sarndal, and M.P. Singh, Eds.), John Wiley and Sons, New York, 1987, pp. 124–137.

31. Hui, S. L., and Berger, J. O. Empirical Bayes estimation of rates in longitudinal studies. J. Am. Stat. Assoc. 78: 753–759 (1983).
32. DerSimonian, R., and Laird, N. M. Meta-analysis in clinical trials. J. Controlled Clin. Trials 7: 177–188 (1986).
33. Wiley, J. A., Herschkorn, S. J., and Padian, N. S. Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penile-vaginal intercourse. Stat. Med. 8: 93–102 (1989).
34. Laird, N. M., and Louis, T. A.. Empirical Bayes confidence intervals based on bootstrap samples (with discussion). J. Am. Stat. Assoc. 82: 739–757 (1987).
35. Colton, T., Freedman, L. S., and Johnson, A. L., Eds. Proceedings of the Workshop on Methods for Longitudinal Data Analysis in Epidemiological and Clinical Studies. Stat. Med. 7(½) (1986).
36. Louis, T. A. General methods for analyzing repeated measures. Stat. Med. 7: 29–45 (1988).
37. Burr, D. On errors-in-variables in binary regression—Berkson Case. J. Am. Stat. Assoc. 83: 739–743 (1988).
38. Fuller, W. A. Measurement Error Models. John Wiley and Sons, New York, 1987.
39. Gail, M. H., Wieand, S., and Piatadosi, S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. Biometrika 71: 431–444 (1984).
40. Mansky, C. R. Identification of binary response models. J. Am. Stat. Assoc. 83: 729–738 (1988).
41. Bailar, J. C. III, Crouch, E. A. C., Shaikh, R., and Spiegelman, D. One-it models of carcinogenesis: conservative or not? Risk Anal. 8: 485–497 (1988).
42. Stiratelli, R., Laird, N., and Ware, J. H. Random effects models for serial observations with binary response. Biometrics 40: 961–972 (1984).
43. Aalen, O. Two examples of modeling heterogeneity in survival analysis. Scand. J. Stat. 14: 19–25 (1987).
44. Hougaard, P. A class of multivariate failure time distrubutions. Biometrika 73: 671–678 (1986).
45. Manton, K. E., Stallard, E., and Vaupel, J. W. Alternative models of mortality risks among the aged. J. Am. Stat. Assoc. 81: 635–644 (1986).
46. Schumacker, M., Olschewski, M., and Schmoor, C. The impact of heterogeneity on the comparsion of survival times. Stat. Med. 6: 773–784 (1987).
47. Trussell, J., and Richards, T. Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. In: Sociological Methodology, Jossey-Bass, San Francisco, CA, 1985, pp. 242–276.
48. Dempster, A. P., Sewlyn, M. R., and Weeks, B. J. Combining historical and randomized controls for assessing trends in proportions. J. Am. Stat. Assoc. 78. 221–227 (1983).
49. Tamura, R. M., and Young, S. S. The incorporation of historical control information in tests of proportions: simulation study of Tarone's procedure. Biometrics 42: 343–349 (1986).
50. Federal Register. Vol. 46, No. 238, p. 60656, 1981.
51. Freedman, D. A., and Zeisel, H. From mouse-to-man: the quantitative assessment of cancer risks (with discussion). Stat. Sci. 3: 3–56 (1988).
52. DuMouchel, W. H., and Harris, J. E. Bayes methods for combining the results of cancer studies in humans and other species (with discussion). J. Am. Statist. Assoc. 78: 293–315 (1983).
53. Laird, N. M. Thyroid cancer risk form exposure to ionizing radiation: a case study in the comparative potency model. Risk Anal. 7: 299–309 (1987).
54. Crouch, E., and Wilson, R. Interspecies comparison of carcinogenic potency. J. Toxicol. Environ. Health 5: 1095–1118 (1979).
55. Amato, D. A., and Lagakos, S. W. Analysis of agreement among findings of pathologists in ED_{01} experiment. J. Natl. Cancer Inst. 80: 919–925 (1988).