Something Specious in the Air? Some Statistical Misconceptions in Aviation Safety Research

ARNOLD BARNETT

Because fatal air crashes are rare, it is often asserted that data about them cannot yield reliable inferences about patterns in air safety. That assertion may overstate, however, both the limitations of small data samples and the sensitivity of analytic outcomes to changes in starting assumptions. The issue is explored by discussing criticisms by a TRB panel and other commentators of two recent air-safety papers cowritten by this author. The goal is not to suggest that monitoring aviation safety can be reduced to studying fatal crashes, but to avoid an unnatural deemphasis on such crashes because of statistical misunderstandings. Beyond aviation, this may interest the broader group of transportation researchers who work with small data samples.

"I have been vilified; I have been crucified; I have even been criticized" (Mayor Richard J. Daley).

Not only do Americans care enormously about aviation safety, but their perceptions on the subject substantially affect their flying behavior. At critical times in spring 1986 and winter 1991, fear of terrorism cut transatlantic air travel by almost 50 percent. Anxiety about sabotage during the Gulf War caused the cancellation of millions of U.S. domestic air trips. And in the first 2 weeks after the 1989 DC-10 crash at Sioux City, Iowa, new bookings on the controversial DC-10 jet may have fallen by more than 33 percent (*1*,p.45).

Such circumstances lend exceptional interest to studies about risk patterns in U.S. commercial aviation, and many researchers have investigated the topic in one way or another. The author and his students studied the topic by analyzing data about fatal air crashes, generally weighting each such event by the proportion of passengers who perish in it. (*2*,p.1045;*3*;*4*,p.1;*5*,p.8;*6*,p.1). Although such fatal crashes are rare, it was argued that many striking phenomena in the data cannot plausibly be dismissed as random fluctuations.

Several papers describing this work have been harshly criticized for both the methodology employed and conclusions reached. Critics have included representatives of TRB, the National Transportation Safety Board (NTSB), FAA, the Air Transport Association, and the academic community. Even the television show *Saturday Night Live* hinted that one of the analyses was a bit preposterous.

No personal offense is taken at such negative assessments; professors wear bullet-proof vests under their academic robes. But the cumulative effect of the criticisms may be to suggest that it is unwise and perhaps irresponsible to perform safety studies that focus on plane crashes, the very events that air travelers fear most. This article takes issue with the critics and defends the "fatal event" approach to analyzing air safety. The goal is not to suggest that monitoring air safety can be reduced to studying fatal crashes; instead the goal is to avoid an unnatural deemphasis on such crashes because of statistical misunderstandings. To the extent that small-sample issues arise elsewhere in transportation research, the arguments in this paper are of wider relevance.

This article is concentrated on two recent manuscripts and the reactions they evoked (*4*,*6*). After a summary of each paper and major objections to the paper's content, attempts to rebut the objections one by one are offered. Other approaches to analyzing air safety are briefly discussed, suggesting that they might not be fully satisfactory.

AIRLINE SAFETY: THE LAST DECADE

Original Analysis

Written in late 1987 and published in early 1989, "Airline Safety: the Last Decade" (*4*) analyzed safety data for 120 airlines from 1977–1986. (The phrase "last decade" refers to the 10 years since the observation period for a previous MIT air-safety study (*2*). Contending that "the greatest fear in aviation is of being killed in a plane crash," it concentrated on statistics about the likelihood of that outcome. It noted that fatal-accident data are not perfect descriptors of system safety (or predictors of future safety) but pointed out that "no serious discussion about aviation risk can be oblivious to the objective trends in actual safety performance."

Most of its calculations were motivated by the question, If a passenger selected one flight completely at random within the set of interest (e.g., U.S. domestic jet flights in 1983), what is the probability that the passenger would be killed in an air crash? It argued that this death-risk-per-flight statistic was a more stable and illuminating measure of hazard than two widely used indicators of mortality risk, namely, deaths per billion passenger miles and deaths per million passengers carried. The last two measures weight each crash solely by the number of passengers killed, without reference to how many passengers were on board. In terms of system safety, however, a crash with perhaps 28 deaths might have very different implications if it reflected a high survival rate on a

Sloan School of Management, Massachusetts Institute of Technology, E53-379, Cambridge, Mass. 02139.

heavily crowded plane rather than a zero survival rate on a lightly loaded one. The death-risk-per-flight indicator avoids such ambiguity by weighting crashes by the percentage of passengers who perished.

Updating the earlier MIT study, the authors prepared mortality-risk estimates for three groups of airlines. These estimates and their counterparts for the earlier observation period are shown in Table 1. The table depicts large (and statistically significant) risk differences within each period among the three airline groups periods. But in a striking pattern of parallel improvement, all three groups cut the death risk by more than 80 percent between the earlier study period and the later one.

The authors also took account of an extraordinary development during the decade of study: the deregulation of U.S. domestic aviation. Before passage of the 1978 Airline Deregulation Act, "fears were expressed that, even though the new airlines brought into existence by the Act would be bound by federal safety rules, they would not match the accomplishments of established carriers with decades of experience" (*4*). Whether such misgivings were borne out by subsequent events was explored.

By focusing on the 8 years from 1979 (the first year of deregulation) to 1986 (the end of the observation), the authors computed mortality risk indicators for both established domestic airlines formed well before 1978 and for new entrant carriers spawned by deregulation. To make the comparisons fair, the authors considered only the 19 airlines among the new entrants that had all-jet fleets. Table 2 presents a summary of the key findings. The findings imply that death risk per flight from 1979 to 1986 was 12.2 times higher on the new jet entrants as on the established carriers. But air crashes were so rare in the study period that all risk estimates were intrinsically imprecise. Before treating the disparity in Table 2 as noteworthy, therefore, the authors performed a test of its statistical significance.

The analysis began with the conservative null hypothesis that new entrants and established carriers were equally safe. Nearly all domestic jet deaths from 1979 to 1986 occurred in five disasters that, on average, killed 95 percent of the passengers. The authors argued that, because the new entrants performed 4.7 percent of U.S. jet takeoffs and landings from 1979 to 1986, an equal-safety hypothesis would assign them a 4.7 percent chance of suffering any jet disaster during the period. Thus, their share of the five domestic jet disasters would be governed by a binomial probability distribution with five "trials" and parameter value of 0.047.

It emerged that the new entrants suffered two of the five crashes (40 percent of the disasters with 4.7 percent of the flights). That outcome does not absolutely prove that the new entrants were less safe than established carriers. But exact binomial calculations reveal that there is only a 1 in 42 chance that, solely because of bad luck, the new carriers would sustain as disproportionate a share of 1979–1986 disasters as they actually did. At the usual 5 percent significance level, therefore, the equal-safety hypothesis would be rejected, and hence

TABLE 1 Death Risk per Flight on Three Airline Groups and Two Successive Periods

Airline Group	1960-75	1976-86
Established U.S. Domestic Carriers	1 in 1.5 million	1 in 11.6 million
First-World Flag Carriers	1 in 430,000	1 in 4.4 million
Second/Third World Flag Carriers	1 in 67,000	1 in 390,000

Note: A nation's flag carrier was defined as its leading international airline. Mortality risk estimates for flag carriers were based only on their scheduled international operations.

TABLE 2 Death Risk for Flight on Two Groups of Domestic Jet Airlines (1979–1986)

Airline Group	Death Risk Per Flight	Percentage of Scheduled Domestic Flights, 1979-86
Established Carriers	1 in 11.8 million	95.3
New Entrants	1 in 870,000	4.7
All U.S. Jet Carriers	1 in 7.4 million	100.0
(All U.S. Jet Carriers, 1971-78)	1 in 2.6 million	100.0

the large discrepancy in Table 2 would be given both practical and statistical significance.

Table 2 also shows that, despite the relatively weak record of the new entrants, overall domestic death risk per jet flight fell from 1 in 2.6 million in 1971–1978 to 1 in 7.4 million in 1979–1986. But if deregulation had not occurred (hence no new entrants), the established carriers would presumably have performed virtually all domestic jet flights from 1979 to 1986. Assuming that deregulation neither worsened nor improved the 1979–1986 record of such carriers, it is suggested (Table 2) that, without deregulation, death risk per domestic jet flight from 1979 to 1986 would have been roughly 1 in 11.8 million (as opposed to the 1 in 7.4 million that actually prevailed). Because the increase was statistically significant, the assertion that overall domestic jet safety improved after 1979 despite deregulation and would have improved more without the policy shift was described as "plausible."

Reactions to Paper

"Airline Safety: The Last Decade" (*4*) attracted considerable attention, some of it positive. The death-risk-per-flight statistic was described as a conceptual advance, and the findings about the safety of established domestic airlines were well received. Attention was given to its calculation that, if a domestic-jet passenger chose one flight at random each day, the passenger would on average travel for 29,000 years before dying in a fatal crash.

The conclusions about airline deregulation, however, engendered a different response. The suggestion that the policy had had an adverse effect on safety was declared unreliable, and the analysis that produced it was portrayed as shallow or self-contradictory.

TRB Panel

Probably the greatest blow to the credibility of the work on deregulation was its sharp rejection by a TRB panel (*7*,Ch.5). The panel reached its negative verdict for three reasons.

1. It took issue with the definition of post-deregulation new entrants and, in particular, with the placement of World Airways in that category. [Because World Airways had amassed "extensive (pre-deregulation) experience in jet charter operations," the panel questioned the inclusion of its 1982 fatal accident in the new-entrant risk calculation.]
2. It noted that the inferences about new-entrant safety were derived from "only three fatal accidents" and hence saw them as subject to great instability.
3. It saw inconsistency in the authors' reasoning about the effects of deregulation:

> If deregulation was somehow responsible for allowing new entrant carriers, who provide about 5 percent of departures, to operate at a higher risk, then deregulation must also be given credit for the almost fivefold reduction in risk of the established carriers, who provide 95 percent of departures. Neither of these conclusions, however, seems plausible, especially in light of (other) studies. (*7*,p.169).

The panel said, "the risk definition and data aggregation techniques used by Barnett and Higgins do not measure the effects of deregulation on safety."

These words appeared in a comprehensive report on airline deregulation that has quite properly earned national distribution and respect. But, as much as the TRB panel conducted intellectually vigorous reviews of published literature, so should its own analysis be subject to spirited inquiry. It can be argued that all three of the panel's criticisms of the authors' work are fully rebuttable, and its rejection of the authors' conclusions should not stand.

There is room for argument about whether World Airways was really a new entrant (Point 1 of TRB). (The issue is whether operating international charter flights provides adequate preparation for scheduled domestic service.) But the debate need not be pursued, because it is irrelevant to the authors' conclusions. As noted, the authors weighted individual crashes by the proportion of passengers killed, and (as made clear) only 1 percent of those on board died when a World Airways jet skidded into Boston Harbor. Because its statistical test of the equal-safety hypothesis considered only the five full-fledged disasters from 1979 to 1986, it did not count the World Airways incident at all. Moreover, excluding World Airways data from the new-entrant calculation increases the estimates of the group's death risk per flight, from 1 in 870,000 during 1979–1986 to 1 in 855,000.

The reference to an unstable risk estimate (Point 2 of TRB) on the basis of only three fatal accidents is legitimate; indeed, to underscore the imprecision of the estimate, the authors calculated a statistical confidence interval for it. In context, however, the panel was tacitly advancing the stronger position that no serious conclusion about new-entrant safety can be drawn from a few fatal crashes. That is an overstatement.

Suppose that Air Scarsdale began jet operations between Westchester County and Hollywood-Burbank airports and that on its first day of service it lost two planes in fatal crashes. Can anyone imagine that passengers would deem it premature to judge Air Scarsdale's safety level based on "only two" disasters or that the carrier would even exist a second day? To put the issue more generally, even a few fatal crashes can be enormously discouraging if one would have expected none under normal conditions.

On the basis of both the disaster rate of established carriers and the number of annual flights performed by new entrants, one would have expected the new entrants to go 50 years before suffering the first disaster. In reality, they experienced two disasters in the first 8 years of service. Such an outcome is not inconceivable under the equal-safety hypothesis, but it is highly improbable. Few statisticians would see the outcome as leaving the hypothesis unscathed.

The final argument (Point 3 of TRB) was that if deregulation was somehow responsible for the relatively poor performance of new entrants, the policy must be given credit for post-1979 safety gains on the established airlines. But an important distinction must be made. It is plausible to tie deregulation to the new entrants' safety record because the end of regulation was responsible for their existence. Yet there is no corresponding requirement to treat deregulation as the main (or only) determinant of recent safety trends on other U.S. carriers. That position is obviously challenged by evidence in Table 1, which demonstrates that however impressive the

recent risk reductions on the new entrants were no greater on a proportional basis than those on international flag carriers wholly unaffected by deregulation.

Other Reactions

In a detailed survey about deregulation's impact on domestic air safety, McKenzie and Womer reviewed the authors' analysis on the topic and found it wanting (*8*). They noted that the inferences about new entrants arose from "exceedingly few accidents," that 16 of the 19 jet carriers formed after deregulation had no fatalities from 1979 to 1986, and that the new entrants may have flown more dangerous domestic routes than did established carriers.

McKenzie and Womer also criticized the authors for dismissing the possibility that deregulation had improved safety on established U.S. airlines. Such progress could have occurred, they argued, because deregulation may have allowed more efficient and profitable operations and produced greater allocations of resources to safety by both government and airlines.

Two of these objections can be dealt with quickly. The response to the comment about "exceedingly few accidents" is the same as the one to the TRB panel's similar statement. The "dangerous routes" argument seems unpersuasive because all three fatal events that befell new entrants during 1979–1986 occurred not at obscure places ignored by established carriers but at the busy airports in Boston, Milwaukee, and Washington.

It is true that 16 of 19 new jet entrants had perfect safety records from 1979 to 1986. But that does not work against any argument advanced by the authors. The aim was to estimate the overall effect of the open-skies policy, and hence what was needed most was the average risk level among the jet "children" of deregulation. Estimating the group average neither implies nor requires that there be no airline-by-airline variation around the mean. At the same time, the "16 in 19" statistic is not strong evidence of heterogeneity in new-entrant safety. Collectively, the 19 carriers averaged one disaster per 850,000 jet flights. But the average new entrant performed only 90,000 flights from 1979 to 1986. Thus, even if the 1-in-850,000 risk level applied equally to all 19 carriers, the vast majority would have had no disasters in the 8 years studied.

As McKenzie and Womer report, the authors' paper quickly discounted the possibility that deregulation had made established carriers safer. But McKenzie and Womer's theoretical arguments on how deregulation increased air safety are not easily reconciled with familiar facts. If better monitoring by federal authorities improved post-deregulation safety on the established airlines, that achievement has gone unrecorded; the government's most conspicuous air-safety activity from 1979 to 1986 was to try to rebuild the air traffic control system after the firing of 11,000 controllers in a 1981 strike. And it seems odd to speak of the financial benefits that deregulation brought to established carriers; all of them lost money, and most suffered economic stresses so great that they were subsequently forced into merger, bankruptcy, or extinction.

On a more methodological level, Oster et al. (*9*) raised a challenge to the authors' analysis of deregulation. Although not referring to that work, these scholars questioned the soundness of the general procedure by which the authors tested the equal-safety hypothesis. Tests like those were "open to question," because they use data that "reflect the *ex post* universe of accident performance of airlines rather than a random sample of such performance" (*9*,p.81).

It is surely true that if one devises a hypothesis upon looking at certain data, there is something circular in using those data for a "test" of that hypothesis. Moreover, if events in the process under study affect the end point and the length of one's period of observation, or both, one gets a biased data sample and not a random one. But when neither of these conditions obtains, gathering data from a particular time interval can readily be construed as a form of random sampling.

Any hypothesis like the equal-safety conjecture is, after all, a description of what will happen in the long run. The events that occur over an innocently chosen short time span are only a random sample of those that would arise if the process were observed indefinitely. And, like the opinions of randomly selected citizens picked in a political poll, these events allow useful but imperfect inferences about the underlying pattern.

The authors choice of 1979–1986 as the observation period was not contaminated by any visible threats to randomness. The first year that domestic skies were deregulated was 1979, and 1986 ended a 10-year period since an earlier air-safety study at MIT.

Another objection to the authors' analysis was never raised: perhaps what had been observed was only a transient effect of deregulation, tied to the fact that new entrants were at the start of the learning curve. That possibility is not outlandish, but Table 1 shows that the higher risk levels of flag carriers in developing countries have proved enduring rather than ephemeral. In any case, if the new entrants had been economically successful, a continuing stream of newer entrants may have flowed to the start of the learning curve.

UNFORTUNATE PATTERN IN U.S. DOMESTIC JET CRASHES

Original Analysis

The death-risk-per-flight statistic was predicated on a completely random choice among flights. But passengers do not select flights at random: they appear in larger numbers on the average 747 aircraft than on the average DC-9 aircraft and in larger numbers on the Wednesday before Thanksgiving than on the third Wednesday in January. Such nonuniformities in demand, however, would not bias the death-risk estimate so long as passengers did not travel disproportionately on hazard-prone flights. The authors of "Airline Safety: The Last Decade" (*4*) took it for granted that this last caveat reflected only a remote possibility.

In 1990 Barnett and Curtis, authors of "An Unfortunate Pattern in U.S. Domestic Jet Crashes" (*6*), set out to dispense with the caveat altogether by trying to document that the domestic jets involved in major crashes were neither unusually large nor unusually crowded. These authors concentrated on established U.S. domestic carriers from 1975 to 1989. (Post-deregulation new entrants were excluded, it was explained, because nearly all of them were out of business by 1989 and thus their undistinguished safety records appeared to be of no continuing relevance.)

The authors defined a crash as being major if it killed at least 20 percent of the passengers. Focusing on the percentage rather than the raw number killed avoided a built-in bias toward large or crowded planes. If, for example, a crash were classified as major only if it took at least 250 lives, the "finding" that all major crashes occurred on wide-body jets would not be illuminating. Under this criterion, if a plane with two passengers crashed and one of them was killed, the crash would be designated as major.

Major crashes occurred from 1975 to 1989 that collectively accounted for 98 percent of domestic jet deaths during this period. The authors tested the following null hypothesis:

> The distribution of passenger loads for the 10 jets in major crashes was statistically indistinguishable from the corresponding distribution for all domestic jet flights from the time span 1975–1989.

Surprisingly, the data dictated the emphatic rejection of the hypothesis. The 10 planes in major crashes averaged nearly twice as many passengers as did other domestic flights from 1975 to 1989. Difference-in-distribution tests that compared the histogram of passenger loads on the 10 ill-fated flights with that of the other 60 million jet flights from 1975 to 1989 reached a strong conclusion: if the crashed flights had truly been a random sample from all domestic jet flights, the probability would be only about 1 in 5,000 that they would have carried as many passengers in total as they actually did. (In statistical parlance, the *p*-value of the observed pattern was 1 in 5,000).

The 10 planes that crashed had an average load factor of 84.7 percent, which was more than 25 percentage points higher than the average of 59.4 percent for all domestic jets over the period. Whereas individual jets exhibit wide variation in load factors, the detailed load-factor distribution renders it extremely unusual to be 25 points above average in a randomly drawn sample of 10 domestic jets. The upshot is that passengers did appear to fly in larger-than-usual numbers on hazard-prone flights. That finding undercuts one of the premises of the death-risk-per-flight statistic (although not the authors' (*4*) general conclusion about deregulation, which does not depend on that statistic).

Reactions to Paper

The paper by Barnett and Curtis evoked much press attention and disapproval. The director of research and engineering for the NTSB criticized the study's methodology and conclusions in interviews with the Chicago Tribune (*10*,p.1) and *Seattle Post-Intelligencer* (*11*,p.81). "When you are dealing with extremely rare events like major crashes," he explained, "you have to be extremely careful about extrapolating information." He made clear that he doubted that the researchers had displayed the requisite prudence.

Others joined in expressing strong objections to the paper. A spokesperson for the Air Transport Association (ATA) declared that the authors showed extremely poor judgment in choosing variables for the study. Noting that "there were months of the year when there were no crashes," she asked, "does that mean it's safer to fly in those months?" A spokesperson for FAA told the *New York Times* (*12*,p.A16) that the pattern discovered "must be a coincidence. We've investigated these crashes," he reported, "and we know their causes."

Such dismissive reactions were perplexing to the authors. The NTSB director had accused the authors of being oblivious to small-sample hazards but the authors devoted a full appendix to the subject in the paper. The authors applied several formal tests to the data, all of which gave full weight to the limited number of events under study. The procedures for drawing conclusions followed widely accepted standards of statistical inference.

A simple comparison offers some perspective on the strength of the authors' finding. If one tossed a coin 10 times and got all heads, one would presumably be highly skeptical that the coin was fair. Although 10 heads in a row from a fair coin is freakish (a 1-in-1,024 chance), it is far less so than picking 10 domestic jets at random and finding as many passengers as were actually aboard the 10 planes that crashed (1 in 5,000).

It is unclear why the ATA spokesperson criticized the authors for extremely poor judgment. The authors did not assume a link between crowding and safety at the outset but, on the contrary, hypothesized its absence. As to the query on whether air travel might be safer in some months than others, the concentration of snowstorms and thunderstorms in certain seasons suggests that the answer could well be yes.

It is not evident how to construe the FAA spokesperson's contention that coincidence explains the findings in the paper. If he means that the outcome reflects nothing more than fluctuations, the 1-in-5,000 probability estimate appears inconsistent with that interpretation. If he means that crowding per se did not cause any crashes, he is restating a point that the authors had made. (The *New York Times* subheadline about the work included the words "no casual link.") The authors had warned, however, that crowding could raise the probability that some other factor could lead to disaster (e.g., improperly deployed flaps) and argued that the risk-crowding correlation could be noteworthy even if causality is absent. If, for example, planes at rush hour are at unusually high risk of airport-area collisions, identifying that pattern gives information to passengers that they might use in deciding when to fly.

NBC's *Saturday Night Live* also got into the spirit of things by summarizing the authors' paper.

> An MIT study has concluded that more people die in planes with more passengers. For instance, in a plane carrying 220 people, 220 people would be killed. As opposed to a plane carrying 15 people, where only 15 people would be killed. A further study reveals that your best chance of survival is to fly in an empty plane (*13*).

RESPONSE TO CRITICISMS

If Barnett and Curtis (*4*) and Barnett and Higgins (*6*) have been subject to excessive criticism, why should that be of general concern? There are a few reasons.

It is said that nature abhors a vacuum, which in this context means that the enormous public interest in aviation safety will inevitably generate statistics on the subject. If direct estimates of the risk of being killed are successfully portrayed as deeply compromised, then proxy risk measures will come to the fore.

But these surrogate statistics may create new problems in the course of circumventing others.

The TRB panel questioned the Barnett and Higgins findings on deregulation in part because of the results of other studies. Only one study that the panel cited contrasted established carriers with new jet entrants (*14*,p.237). That paper noted that compared with established airlines, the new entrants spent larger portions of their operating budgets on maintenance and had slightly fewer accidents per million departures. In effect, the study treated maintenance expenditures and overall accident rates as superior proxies for "true" mortality risk than was the observed mortality risk over the same period.

But were those measures really superior? Maintenance's share of the budget seems an ambiguous measure of safety: the new airlines may have operated older fleets, been less able to take advantage of economies of scale, or spent maintenance money less wisely than their established counterparts. And if higher expenditures on maintenance mean lower ones on other essentials (e.g., pilot training), then safety could suffer rather than gain.

In emphasizing overall accident rates, the study (*10*) cited by TRB was following a common practice in air safety research. That circumstance lends particular importance to two events from the 1980s. The first involved an Aloha Airlines Boeing 737, which suffered an in-flight structural failure that practically destroyed the upper half of its fuselage. A flight attendant was blown out of the crippled plane, but the pilot managed to land it with no passenger injuries. The second took place on an Air Canada Boeing 767 that, because of a misunderstanding about whether its kerosene requirement was expressed in pounds or kilograms, literally ran out of fuel in mid-air. The pilots brought it down safely to an abandoned airstrip in Manitoba. Although the plane was damaged in the highly irregular landing, no passengers were hurt.

Both of these events meet a broad definition of accidents. But is it irrelevant that extraordinarily skilled cockpit crews saved all the passengers from airborne crises that could easily have killed them all. Arguably, the consequences of an accident say more about the safety of an airline's operation than does the existence of the accident. Yet such consequences get no weight at all in overall accident-rate statistics.

More generally, proxy measures for the death risk of flying may avoid such unpopular activities as inferences from small data samples. But the proxies typically entail questionable assumptions and blurring of salient distinctions. It is not obvious that they are more illuminating than direct measures of mortality risks. Some cures, as the saying goes, are worse than the diseases.

But, as a practical matter, do any conclusions drawn about the safety of air travel depend heavily on the way it is measured? Clearly, the answer can be yes as evidenced by the TRB panel's sharp contrast between the authors' findings and those from another study. Effects that are large and statistically significant under one measure may be nonexistent under another. The selection of a safety index, therefore, is a matter of more than aesthetic interest.

And, of course, the conclusions that one draws about prevailing safety patterns affect one's perceptions about how to reduce risk. The authors received several calls about the paper by Barnett and Curtis (*6*) from a senior captain at one of America's leading airlines. He reported that pilots and copilots alternate takeoffs at his carrier but that in inclement weather the pilot always takes the controls. He wants to adopt the bad-weather rule on fully loaded long-distance flights because of a reduced margin of error in dealing with takeoff emergencies. The authors' statistical findings, he reported, would be helpful to him in making his case.

Perhaps his airline and others will adopt the policy change he is seeking, and perhaps over the next quarter century, one jet takeoff crash will thereby be averted. That possibility alone suggests that the authors acted properly in reporting the pattern that they had observed. It also pointedly suggests that dismissing the finding out of hand might not be a risk-free option.

CONCLUDING REMARKS

Because fatal air crashes are rare, analyzing data about them means working with small samples. Small-sample data are volatile, and even apparently stark patterns within them may be nothing more than meaningless fluctuations. If the fatal crashes are partitioned into categories, slight changes in the classification rules might substantially alter cross-category differences.

Serious researchers recognize these hazards. But they also recognize that whereas small data samples are not inevitably useful, neither are they inevitably useless. By means of formal tests of statistical significance, calculation of confidence intervals for key parameters, and sensitivity analyses to see whether the findings depend substantially on particular decisions (e.g., on whether to classify World Airways as a new entrant), the researchers can realistically assess whether particular results are too imprecise to be credible. Statistically minded investigators understand that if a pattern is sufficiently extreme, a clear signal can be transmitted by even a small data sample.

Official reactions to studies about U.S. air disasters often appear defensive. But such defensiveness is misplaced: such studies do not disadvantage U.S. aviation but, on the contrary, constitute the most effective means of upholding the claim that established U.S. carriers are the world's safest airlines. It would be ironic if bodies like FAA, ATA, and NTSB succeeded in discrediting the form of analysis that tells us exactly what they are: perhaps the most successful organizations devoted to safety in the history of the world.

ACKNOWLEDGMENTS

The author benefited greatly from perceptive comments by Steve Cohen, Richard Golaszewski, Jesse Goranson, Robert E. Machol, Michael Peterson, and Amedeo Odoni, none of whom should be assumed to hold any given viewpoint expressed in the paper.

REFERENCES

1. Barnett, A., J. Menhigetti, and M. Prete. The Market Response to the Sioux City DC-10 Crash. *Risk Analysis*, Vol. 12, No. 1, March 1992.

2. Barnett, A., M. Abraham, and V. Schimmel. Airline Safety: Some Empirical Findings. *Management Science*, Vol. 25, No. 11, Nov. 1979.
3. Barnett, A. See Lightning? Close Airports. *The New York Times*, June 26, 1986 (op-ed page).
4. Barnett, A., and M. K. Higgins. Airline Safety: The Last Decade. *Management Science*, Vol. 35, No. 1, Jan. 1989.
5. Barnett, A., Air Safety: End of the Golden Age? *Chance: New Directions for Computing and Statistics*, Vol. 2, No. 3, Summer 1990.
6. Barnett, A., and T. Curtis. An Unfortunate Pattern in U.S. Domestic Jet Accidents. *Flight Safety Digest*, Vol. 10, No. 10, Oct. 1991.
7. *Special Report 230: Winds of Change: Domestic Air Transport Since Deregulation*, TRB, National Research Council, Washington, D.C., 1991.
8. McKenzie, R., and N. Womer. *The Impact of the Airline Deregulation Process on Air-Travel Safety*. Working paper 143. Center for the Study of American Business, Washington University, St. Louis, Mo., Sept. 1991.
9. Oster, C. V., J. S. Strong, and C. K. Zohn. *Why Airlines Crash: Aviation Safety in a Changing World*. Oxford University Press, New York, 1992.
10. *Chicago Tribune*. Oct. 19, 1991, p. 1.
11. *Seattle Post-Intelligencer*. Oct. 17, 1991, p. 1.
12. *New York Times*. Nov. 1, 1991, p. A16.
13. *Saturday Night Live*. NBC. Nov. 2, 1991.
14. Kanifani, A., and T. Keeler. New Entrants and Safety: Some Statistical Evidence on the Effects of Airline Deregulation. *Proc., Transportation Deregulation and Safety*, Transportation Center, Northwestern University, Evanston, Ill., 1988.

DISCUSSION

TRB Task Force on Statistical Methods in Transportation (A3T51)

REVIEWER 1

Barnett and Higgins, in their work "Airline Safety: The Last Decade" (*1*), propose the Q-statistic and use it as the measure for reporting risk (and relative risk when comparing carrier groups). However, the article makes no real inquiry into the sampling distribution of Q, and as shall be discussed here, it appears to this reviewer that the authors underestimate the variability inherent in the risk elements.

All the hypothesis tests in the authors' work use binomial tests (randomization tests) on the frequency of "disasters." These tests are conditional upon a classification of fatal accidents as disasters and nondisasters. The classification occurs after the fact and any variability present in the proportion is disregarded. Overall the authors report 12 fatal accidents during the period 1977 to 1986, of which six are classified as disasters. Of the nine fatal accidents from 1977 to 1986 reported by the authors for trunkline carriers, four are classified as disasters, and for the period 1979 to 1986, three of (presumably) eight are also classified as disasters. For new entrant airlines, two of three accidents are classified as disasters. The disaster classification rates are 0.375 and 0.667, respectively, for established carriers and new entrants, and the hypothesis test in Section 7 of the authors' work is contingent on those observed rates.

There are certainly reasons to anticipate that differences in crew training and experience, equipage, and other factors might affect the expected survival rate in a given accident or in a population of accidents. On the other hand, even given the specific set of circumstances that attends a particular accident, the intuition of this reviewer is that mortality in the accident is still very much a matter of chance. The next three accidents are very much a matter of chance. The next three accidents occurring to new entrants, from 1987 on, for example, might well produce no disasters, or one, or three. The number of disasters arising from a specified number of fatal accidents is arguably binomial with an unknown parameter value that may depend on the type of airline or environmental risk factors. The variability of this binomial type of outcome should be incorporated when making inferences regarding disaster incidence. In addition, the equal safety hypothesis of the authors should include the supposition that the mean proportional mortality is the same for established carriers and new entrants, and (in this regard there is a small-sample problem) the supposition would not be rejected.

Similarly, the weighting factors used in constructing Q are random variables, embodying the "fluctuations . . . in the survival rate per incident," and in contrast to the dismissal of such fluctuations by the authors as second-order effects, this reviewer believes that they make a nonnegligible contribution to the variability of Q when the number of accidents involved in computing Q is small. Thus it is plausible, but not altogether clear, that the value 12.2 is the maximum likelihood estimate of the new entrants' risk multiplier (*1*,p.15), and if so, work remains to clarify the form of the likelihood. The 10 percent confidence range for the risk multiplier should also probably be wider than as calculated by the authors.

The Q-statistic does have some appeal as an estimator of death risk. It is apparently unbiased. It recognizes the clumping of mortality risk by accident and leads to a sample size (number of departures) that is more meaningful than the number of passenger-departures. However, a more satisfying approach to this reviewer would be to use a hierarchical model, assuming that fatal accidents arise as a binomial or Poisson random variable, and use a second random variate—beta, for example—to model the proportion of deaths, conditional on occurrence of an accident. Such an approach would seek to extract information from all the accident data, at least fatal accident data, instead of ignoring the low fatality accidents in making inferences about safety.

Despite the expressed reservations, it is still likely that the new entrants and established carriers exhibit a statistically significant difference in risk, and the authors are justified in asking that this difference be considered seriously, despite the small number of accidents overall. The reservations are not strictly speaking small sample issues, although the concerns are magnified by the small number of accidents. Rather, they are concerned with modeling technique, in the sense of identifying the proper sampling frame and assessing sources of variability in the sampling frame.

The propositions that are considered in Section 8 of the authors' work (*1*) appear to this reviewer to drift beyond the sampling framework within which the data were collected. Even if one shows that the new entrants have a death risk that exceeds the risk of established carriers, the contention that deregulation "raised by roughly 60% the average risk per flight for domestic jet travel" (*1*,p.16) is based on a causal proposition that is neither proven nor disproven by the data

and previous analyses. In calculating the overall postderegulation death risk, the authors include six local jet carriers formed before deregulation: Aircal, Alaska, Aloha, Hawaiian, PSA, and Southwest. Together these carriers had over 600,000 revenue departures in 1986 (compare with 1.7 million departures on the new entrant airlines, 1979–1986, as given by the authors). To make an assessment of deregulation's impacts, it would be desirable to consider also the experience of these and similar carriers before and after deregulation. Would these former intrastate carriers be naturally aggregated with the new entrants (to the extent that they also experienced rapid growth) or with the established truck carriers? The point to be taken from the third criticism of the TRB panel discussed by Barnett in his paper is that discipline and consistency are necessary in the application of the control variable (i.e., deregulation). The authors make suggestive findings about the death risk of new entrants, and these findings are worthy of serious consideration, in spite of the reservations about sampling distribution and the small number of accidents. However, the manner of presentation by the authors, in the abstract and the final remarks, places relatively emphatic statements about risk on the marquee. The reservations to which Barnett reacts in his paper in part reflect reasonable concerns that the strength of the evidence in the data and the thoroughness of the analyses in the authors' work (*1*) do not yet warrant a determining influence on public policy. Barnett in his paper does not offer enough that is new to overrule those concerns.

Twenty-two fatal accidents are given for 14 C.F.R. Part 121 scheduled passenger operations in the years 1975 to 1986 (*2*,Table5.4). This number excludes accidents involving weather turbulence, sabotage, or a nonoperational event (ramp activities). By further excluding two mid-air collisions and three accidents involving air traffic control or maintenance personnel, 17 fatal accidents are tabulated. A total of 19 fatal accidents between 1977 and 1989 for a subset of the trunkline carriers studied by the authors are given by Neyman and Pearson (*3*,Table5.8). The authors' consider nine fatal accidents for the trunkline carriers in the years 1977 to 1986, and three for the new entrants. Discrepancies of this sort are common in my experience with data on air transportation and may result from different definitions, reporting methods, or inclusion and exclusion criteria.

REVIEWER 2

In his paper, the author begins by declaring that his goal is "to avoid an unnatural deemphasis on [fatal] crashes because of statistical misunderstandings." After a review of his paper, however, it would appear that the author is guilty of an overemphasis on fatal crashes by statistically "stretching" a very small amount of information. Having considerable experience in the analysis and inference of small, rare probability events (traffic accident data and cancer clinical trials), this reviewer feels that too much has been said about too little in the paper and previous articles referred to in this manuscript.

Whereas it is true that one should not uniformly dismiss information based on small sample sizes, by the same token, one should not exaggerate the potential meanings of conjectures based on this information. Such statements as "the overall domestic jet safety . . . would have improved more without the [deregulation] policy shift" based on only three fatal crashes are indeed a "stretch" of statistical inference. That the statistical tests of hypotheses that there is no difference in the safety of established carriers before deregulation and new entrants are rejected at some significance level does not mean that (a) there is, in fact, a true difference in their safety or (b) this difference is caused by deregulation. Two basic flaws are inherent in such conclusions.

- When testing a statistical hypothesis on the basis of one given set of data, the conclusion is simply that we fail to have sufficient evidence to not reject the hypothesis. (Grammatically it would be more appealing to avoid a double negative but this would require the use of the word "accept." By the same token, the result of a given data set never justifies acceptance of a hypothesis, merely failure to reject.). This does not mean that the hypothesis (equality of safety) is false.
- A causal relationship cannot be established by a single study, especially a retrospective, noncohort study (*4*).

By definition, the analyses in this and related manuscripts are based on what is called a case-control, retrospective study in epidemiology. Studies that try to relate the effects of an exposure factor, such as cellular telephones, to rare diseases, such as malignant brain tumors, use this approach and are often guilty of extrapolating more from the data than is justified. In this example, the study is retrospective because it is based on events that occurred before the analysis and is considered a case-control study because the case of new entering flights is being compared with a "control" (established carriers). Although the established carriers are not a control by standard definition, when comparing two groups, one group is termed the "control" in the epidemiological vernacular. At any rate, a common misinterpretation of these results is to interpret a "relative risk" and infer cause and effect. The author has done both in interpreting the safety of the two airline groups. "When both the supposed cause and effect of interest are rare in the general population, the standard retrospective methods often lack sufficient statistical power to evaluate the association of these factors" (*2*). The *Encyclopedia of Statistical Sciences* also states that in such a study design, the rates of the outcomes (fatalities) within groups (existing and new entrants) cannot be estimated with any reliability (*2*).

The conclusions drawn from this study of rare events (fatal plane crashes), which compares two groups on the basis of extremely low occurrences of these events, are equivalent to comparing two cancer treatments and making a decision as to which treatment is best on the basis of a very few subjects. In such cases, information about the few but meaningful subjects should not be ignored but neither should a decisive conclusion be drawn using inferential statistical methods that require large amounts of information. In those situations, case studies should be relied on and expert knowledge used to try to formulate conclusions to benefit the population. And so it must be with airline crashes. Every crash must be studied in detail and general conclusions drawn on the basis of any observable patterns. Criticizing studies that are based on small numbers is not saying that small numbers are not meaningful but that the numerical assessment of the implications of such studies must be combined with good science and not based

purely on statistical probabilities. In the words of Neyman,

> [Statistical] tests themselves give no final verdict, but as tools help the worker to form his final decision. . . . What is of chief importance in order that a sound judgement be formed is that the method adopted, its scope and its limitations, should be clearly understood. . . . (*3*)

REFERENCES

1. Barnett, A., and M. K. Higgins. Airline Safety: The Last Decade. *Management Science*, Vol. 35, No. 1, Jan. 1989.
2. *Encyclopedia and Statistical Sciences*. S. Kotz and N. L. Johnson, eds.) John Wiley and Sons Inc., Vol. 8, 1988, pp. 122–123.
3. Neyman, J., and E. S. Pearson. *Joint Statistical Papers*. University of California Press, Berkeley, 1967.
4. *Air Carrier Traffic Statistics Monthly*. U.S. Department of Transportation, Research and Special Programs Administration, Center for Transportation Information, Transportation Systems Center, Kendall Square, Cambridge, Mass.

AUTHOR'S CLOSURE

I thank the discussants for their thoughtful comments about my paper. Because their remarks concentrated heavily on my airline deregulation work with Higgins "Airline Safety: The Last Decade," I focus my response on that particular data analysis.

Let me begin with some background. It was widely asserted in the late 1980s that, because U.S. air travel was statistically safer after deregulation than before, the policy shift could not have had an adverse effect on passenger safety. But Higgins and I argued that the germane comparison was not between safety levels in the 1980s and those in the 1970s, but between risks in the 1980s and those that would have prevailed at that time if deregulation had not occurred. We performed several calculations to facilitate the latter comparison.

We emphasized those scheduled domestic jet flights from 1979 to 1986 that had resulted in passenger fatalities. We put the accident into two groups: those on which at least half the passengers had died (the disasters) and those on which the majority survived. All events in both categories entered our risk estimates for air travelers; in testing particular patterns for statistical significance, however, we considered only the disasters.

The first reviewer is troubled that we only partitioned fatal events as we did after the fact. In reality, we were following a convention from an earlier paper written before deregulation (*1*,p.1045). Still, the partitioning rule may seem odd: why should a crash that kills 51 percent of the passengers be treated differently from another that kills 49 percent?

Following the pattern of earlier years, nine fatal events from 1979 to 1986 on scheduled U.S. domestic jet flights emerged as heavily polarized between those in which almost no one survived and those in which almost everyone did (Table 3). Five disasters in the table caused more than 99 percent of domestic jet deaths over the 8-year period. As a practical matter, therefore, disaster risk and total risk are almost the same.

With fewer than 5 percent of domestic jet flights from 1979 to 1986, the new entrants suffered 40 percent of the disasters (2 out of 5). The death risk per flight was more than 12 times that of the established carriers. Reviewer 1 suggests that rather than compute an overall statistic, we consider passenger risk in two stages:

1. What is the probability that a randomly chosen flight results in any passenger fatalities?
2. Given that there were such fatalities on a flight, what is the probability that a randomly chosen passenger aboard was killed?

It is suggested that the answer to the second question was roughly 58 percent for the established carriers and 65 percent for the new entrants, corresponding to survival rates of 42 and 35 percent, respectively (Table 3).

These survival rates for the two airline groups are not very far apart. But this similarity does not render the factor-of-12

TABLE 3 Fatalities on Two Groups of Domestic Jet Airlines (1979–1986)

	Airline	Date	Percentage of Passengers Killed
ESTABLISHED CARRIERS:	(1) American	6/79	100 (%)
	(2) Pan Am	7/82	100
	(3) Delta	8/85	83
	(4) USAir	1/79	5
	(5) Republic *	1/83	3
	OVERALL AVERAGE		58
NEW ENTRANTS:	(6) Midwest Express	9/85	100
	(7) Air Florida	2/82	95
	(8) World	2/82	1
	OVERALL AVERAGE		65

* Republic Airlines, formed from the merger of Southern, Hughes Airwest, and North Central, subsequently became part of Northwest.

statistic misleading. We must, after all, also consider the reviewer's first question, the answer to which reveals that new entrants were far likelier than other carriers to suffer fatal events. Delving into the reasons for an overall risk disparity is surely sensible; although we must be careful lest the complexity of the inquiry obscure the magnitude of the effect it is trying to explain.

Reviewer 1 thinks we went too far in suggesting that deregulation raised the risk of domestic jet travel by about 60 percent. From 1979 to 1986, the death risk per flight on established U.S. airlines was 1 in 11.8 million. Yet because of the weaker record of the new entrants, the overall risk level for U.S. domestic jet travel was 1 in 7.4 million. This second statistic is 1.6 times (60 percent higher than) the first. Higgins and I acknowledged that the risk multiplier of 1.6 was subject to great instability: the confidence interval for the multiplier ranged from 1.03 to 5.02. But, as a first approximation for the effect of deregulation on death risk, 1.6 is easier to defend than most other candidates.

Reviewer 1 also wonders why we grouped six regional airlines—Aircal, Alaska, Aloha, Hawaiian, PSA, and Southwest—with giants like United and Delta rather than with new entrants much closer to their size. A critical reason was that years before deregulation PSA and Aircal were the main providers of California's massive intercity jet service. The same is true about Southwest in Texas, Alaska Airlines in Alaska, and Aloha and Hawaiian Airlines in Hawaii. To treat such airlines as "children" of deregulation, therefore, would seem historically inaccurate.

Despite qualms, Reviewer 1 concedes that the risk disparity we reported was "likely" of statistical significance and hence that we were justified in calling attention to the disparity. The negative summary judgment of Reviewer 1 of our work appears to reflect the view that, even if the new entrants were less safe than other airlines, one cannot say that deregulation was responsible for the difference. It is true that correlation does not imply causality; in this instance, however, one might consider deregulation a causal factor because without it, the new entrants presumably would not have come into being.

We have always acknowledged the point of Reviewer 2 that one cannot absolutely prove with statistics that new entrant carriers were intrisicially inferior in safety. (We did not, as Reviewer 2 contends, say that air safety improved recently *despite* deregulation; we made the weaker statement that it was *plausible* to interpret the data that way.) We certainly agree that every air crash should be carefully scrutinized on its own. But such scrutiny does not make the analysis of groups of crashes superfluous; patterns that emerge clearly from group study could well go undetected when each crash is studied in isolation.

Reviewer 2 implicitly compares our work to some unspecified small-sample studies about cancer and automobile accidents. I do not doubt that some people have said more than they should have on the basis of small samples (much as we would have done if we had highlighted the factor-of-12 outcome without considering its statistical significance). But the reviewer's argument appears to boil down to guilt by association, which seems especially unfortunate because the point we were stressing was "not that small samples are inevitably useful, [but] that they are not inevitably useless."

Once when the *Washington Post* was accused of printing an inaccurate news report, it offered the succinct response "we stand by our story." Higgins and I respect and thank the reviewers, but, having considered their reservations about our work, we too stand by our story.

REFERENCE

1. Barnett, A., M. Abraham, and V. Schimmel. Airline Safety: Some Empirical Findings. *Management Science*, Vol. 25, No. 11, Nov. 1979.

Publication of this paper sponsored by Task Force on Statistical Methods in Transportation.