Methods to Identify Pattern Case Failures from I/M Data

United States
Environmental Protection
Agency

Methods to Identify Pattern Case Failures from I/M Data

Office of Mobile Source Emissions
U.S. Environmental Protection Agency

Prepared for EPA by
Energy and Environmental Analysis, Inc.
EPA Contract No. 68-03-1865
Work Assignment No. 20

United States
Environmental Protection
Agency

EPA-420-R-86-001
September 1986

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

EPA terms groups of vehicles that fail the I/M test procedure at an "unusually high rate" as pattern case failures. These groups may fail a particular type of I/M test, or more generally, fail different types of short tests for a variety of reasons The reasons range from design defects common to the particular group, to an emission control system component failure at excessive rates causing vehicles in that group to fail. These "pattern case" failures cause difficulties in I/M programs as such vehicles may not be easily repairable, or it may be appropriate to modify the test procedure for some vehicles. In other cases, it may require EPA to force manufacturers to recall these vehicles for modification.

EPA has traditionally relied on information supplied by individual I/M programs, individual car owners, or in some instances, manufacturers to identify "pattern case" failures. Under a previous work assignment for the EPA, EEA obtained data from three I/M programs and calculated failure rates at the certification engine family level and at several distinct cutpoints The failure rates were utilized by EPA to identify engine families that were potential pattern case failures The objective of this work assignment was to investigate methods to identify pattern case failures using I/M data on a routine basis and to

- Minimize the complexity and time required to obtain such data
- Enhance statistical methods to better resolve pattern cases.

Accordingly, EEA organized the work effort into three separate areas The first area is the data availability, where EEA investigated the quality, quantity and types of data available from I/M programs on a routine basis. The second area is the processing requirement to calculate failure rates by engine family which to some extent depends on the type and cleanliness of data supplied by the states The third area encompasses the statistical tools required to identify pattern case failures, given sample size and observed failure rates We have assumed that EPA is interested in detecting pattern case failures in vehicles that are model year 1981 or newer, since these vehicles are covered by the "207(b)" Emission Warranty Furthermore, our analysis is restricted to light-duty vehicles (LDV) and light-duty trucks (LDT)

This report is organized as follows. Section 2 discusses our findings on the type, quality, quantity and availability of data from seven I/M programs The findings are based on contacts with I/M program managers in seven locations. Section 3 details the processing requirements, starting from raw data as provided by individual I/M programs, to the final product of computed failure rates by engine family Section 4, prepared in conjunction with a subcontractor - Analysis and Simulation, Inc. - provides a range of statistical tools required for identification of so-called "pattern case" failures. Although EPA used the χ^2 test, we believe that more sophisticated methods are required for the analysis Section 5 summarizes our findings and recommends analyses we believe would be of greatest value to EPA.

2. CENTRALIZED STATE AUTOMOTIVE VEHICLE EMISSION INSPECTION/MAINTENANCE PROGRAMS SURVEY

2.1 OVERVIEW

The difficulties associated with the rapid identification of pattern case failures are due to the quality, quantity and availability of test data for analysis. The choice of inspection/maintenance data to analyze pattern case failures must seek to maximize the availability of an adequate sample of data that is relatively error free, contains all of the variables of interest and does not require inordinate delays. Previous analysis for the EPA has shown that failure rates are sensitive to test procedures, and potentially to climatic variables. It is in EPA's interest to obtain data from several different inspection/maintenance programs that represent different test procedures and are geographically dispersed. In fact, failure rate differences arising from test procedural differences or climatic/geographic differences may be quantifiable if there is adequate data.

The requirement for large quantities of relatively clean and unbiased data containing the Vehicle Identification Number (VIN) for each vehicle tested resulted in narrowing the scope of our effort to encompass only centralized I/M program data; analysis of data from decentralized I/M programs have shown that much of the data is suspect. In this analysis, we examined a variety of different centralized I/M programs that represent the range of diversity in location, test procedures and data handling procedures. EPA has also been interested in determining pattern case failures for California vehicles; this, of course, required that we examine the California program even though the program is not centralized The analysis of data from California is considered separately in this section.

A telephone survey was conducted to determine individual characteristics of and differences between the vehicle inspection/maintenance programs of seven states which use centralized (i e., state or contractor operated) inspection facilities. The objective of the survey was to determine which state program(s) produce emissions test data which require little or no pre-processing cleanup or editing, and are available on a frequent (monthly or quarterly) and timely basis. Individuals surveyed were state I/M program engineers and technicians or private contractors involved in the day-to-day operations of the programs The seven programs selected were from diverse geographic regions of the United States: Northeast, Middle Atlantic, Southeast, Midwest, Southwest and Far West. The surveyed states include - Arizona, Connecticut, Illinois, Kentucky, Maryland, Washington and Wisconsin. As stated above, California is also considered but is separate from the analysis of centralized programs.

This section discusses the general testing requirements, sample characteristics, data recording and availability, and special features of each of the seven programs surveyed

2.2 TESTING REQUIREMENTS

Although all state I/M programs ostensibly test light-duty vehicles on the idle test, there appear to be considerable variations in the definition of "light-duty", distinctions between cars and trucks, the actual test procedure used, and the pass/fail requirements. EPA definitions classify all cars as light-duty vehicles and all trucks up to 8,500 lb GVW as light-duty trucks (since 1979) Our survey recorded general confusion regarding the 8,500 lb cutpoint, with some states covering vehicles only up to 6,000 lb and others up to 10,000 lb GVW To the extent EPA is interested in engine families and in light-duty trucks between 6,000 and 8,500 lb GVW, there could be problems in obtaining good data Test methods vary primarily in the preconditioning requirement, although some states now have pass/fail criteria on emissions at high idle or loaded mode tests Cutpoint distinctions between cars and light-trucks are more rare

The survey results are summarized in Table 2-1. The first column in Table 2-1 lists the vehicles eligible for testing under the state requirements. The second column is included to clarify the definition of "light duty" as it pertains to vehicles eligible for testing (henceforth, "vehicles" will refer to cars and trucks, unless otherwise noted) All of the programs test all registered vehicles in the light duty category covering at least the twelve most recent model years, which represent approximately ninety percent of the in-use cars and trucks. Both Illinois and Wisconsin test vehicles up to 8,000 lb (rather than 8,500 lb) GVW, but this distinction is based on registered GVW, which may not be consistent with actual GVW. Washington and Maryland classify vehicles only to 6,000 lb GVW as light-duty while Connecticut tests vehicles to 10,000 lb registered GVW. Complete capture of LDT's between 6,000 and 8,500 GVW is a potential problem with this variation.

The four types of emission tests for gasoline powered light duty vehicles are listed in the third column of Table 2-1 The tests are: 1) Tl - non-preconditioned idle, 2) T2 - loaded mode cruise on a dynamometer, 3) T3 - final idle after preconditioning either at 2500 rpm idle or loaded mode and 4) T4 - 2500 rpm no-load idle. Arizona and Connecticut are unique as they use T3 test only for vehicles which failed Tl test. Arizona, Connecticut, and Wisconsin precondition using a loaded mode cruise on a dynamometer, while the other programs precondition at idle Knowing the tests performed will allow for a comparison of failure rate patterns between different testing sequences to gauge the effect of alternate preconditioning and testing procedures

TABLE 2-1 GENERAL TESTING REQUIREMENTS

State	Vehicles Included	"Light Duty" Definitions	Test Sequence For Gas LDV[a]	Standards/Tests-MYR 1981 and Newer LDV
AZ	All gasoline and diesel vehicles, 1972 and newer[c]	0-8,500 lbs	T1/T2/T3	207(b)[b]/T1 (T3 for failed vehicles)
CT	All gasoline powered cars and trucks, 1968 and newer	0-8,500 lbs	T1/T2/T3	207(b)/T1 (T3 for failed vehicles)
IL	All gasoline powered cars and trucks, 1968 and newer	0-8,000 lbs	T1/T4/T3	207(b)/T4, T3
KY	All gasoline and diesel powered vehicles, all years	0-8,500 lbs	T4/T3	207(b)/T3
MD	All gasoline powered cars and trucks, last 12 years	0-6,000 lbs	T4/T3	207(b)/T3
WA	All gasoline powered cars and trucks, last 14 years	0-6,000 lbs	T4/T3	1 5(2.0)/300/T3
WI	All gasoline powered cars and trucks, last 15 years	0-8,000 lbs	T2/T3	207(b)/T3 (4.0/400 for LDT's 1981-86)

[a] T1 = First Idle, T2 = Loaded Cruise; T3 = Final Idle. T4 = 2500 rpm preconditioning.

[b] CO = 1.2%, HC = 220 ppm.

[c] Will test all 1967+ MYR gas and diesel starting 1/87.

The fourth column of Table 2-1 lists the standards for model years 1981 and newer light duty vehicles. It can be seen that all the surveyed I/M programs except Washington use the U.S. EPA suggested "207(b)" standards of 1.2 percent carbon monoxide (CO) and 220 parts per million hydrocarbon (220 ppm HC) This simplifies the task of comparing the failure rates for all 1981 and newer light duty vehicles across test procedures and states. Illinois is unique in this group in having 2500 rpm idle standards.

2.3 SAMPLE CHARACTERISTICS

It is preferable to have a large clean sample of data for the analysis of pattern case failures. The cleanliness of the sample is affected by retest/multiple test data and appearance of vehicles in the emissions data that are difficult to track or can cause confusion.

Table 2-2 summarizes the characteristics of interest in each I/M program data base Columns one and two pertain to all vehicles (light, medium and heavy duty as applicable) tested, while the third column is specific to light duty vehicles The first column lists the total number of vehicles tested each month The greater the sample size, the more significant any failure patterns detected will be. Sample sizes ranged from 31,000 vehicles per month in Kentucky to 210,000 per month in Illinois for vehicles of all model years.

The overall average failure rates (in percent) are listed in the second column The failure rates are affected by the types and ages of vehicles subject to testing as well as the pass/fail standards in effect. As these failure rates include all classes of vehicles (except for Maryland), they can be used as a general guideline for expected failure rates. EPA may wish to focus analysis towards states reporting higher than average failure rates.

TABLE 2-2 SAMPLE CHARACTERISTICS

State	Monthly Sample Size	Reported Average Failure Rate (%)	Test Requirements
AZ	110,000	81 - 11.9 82 - 8 4 83 - 4 9 84 - 3 0	1. New cars. first anniversary of sale 2 Migrant out-of-state prior to registration 3. Change of ownership A) Sold by dealers prior to sale B) Sold by individual new owner registration renewal (will change to sale 1/87)
CT	133,000	81 - 6 04 82 - 3.91 83 - 2.45 84 - 2 21	1. New cars: first anniversary of sale 2. Migrant out-of-state: prior to registration 3. Change of ownership. VIR renewal
IL	210,000	N/A	1. New cars first anniversary of sale 2. Migrant out-of-state. first anniversary of migration 3. Change of ownership registration renewal
KY	31,000	81 - 11 05 82 - 7 14 83 - 3.09 84 - 2 43	1. New Cars first anniversary of sale 2. Migrant out-of-state registration renewal 3. Change of ownership registration renewal

TABLE 2-2 SAMPLE CHARACTERISTICS (cont'd)

State	Monthly Sample Size	Reported Average Failure Rate (%)	Test Requirement
MD	133,000	81-84 6.3	1 New cars: first anniversary of sale 2. Migrant out-of-state: prior to registration 3. Change of ownership. VIR renewal
WA	50,000	81 - 5 86 82 - 4 42 83 - 3.49 84 - 5.58	1 New cars first anniversary of sale 2. Migrant out-of-state: prior to registration 3. Change of ownership registration renewal
WI	135,000	81 - 11.9 82 - 8 4 83 - 4.9 84 - 3.0	1. New cars. registration renewal (minimum 90 days) 2 Migrant out-of-state: registration renewal 3. Change of ownership: registration renewal

The third column of Table 2-2 summarizes the points at which three classes of vehicles enter the sample data bases The three classes of vehicles for which first test requirements are listed are new cars bought in state, cars being registered in state for the first time (new cars bought out of state and used cars migrating to the state), and used cars that have had a change of ownership. All of the surveyed states waive testing requirements for new cars either until the registration renewal date (no minimum grace period in Kentucky, 90 day grace period in Wisconsin) or until the car is one year old These practices limit the number of new cars, which presumably have a very low failure rate, present in the sample data bases. Migrant out-of-state cars are usually tested prior to first in-state registration, except in Illinois where they are given a one year grace period, and in Kentucky and Washington where they are tested at registration renewal. Used vehicles are often tested at the new owner's registration renewal. Arizona tests used cars sold by dealers prior to sale. Preregistration testing leads to the possibility that the same vehicle will be tested more than once in a twelve month period, thereby biasing failure rate patterns. Moreover, testing of unregistered vehicles results in data with blank fields for VIN number (or nonsense numbers) and license plate Cleaning requirements are therefore higher when such vehicles are present. Connecticut, Illinois and Maryland test vehicles only when the Vehicle Inspection Report assigned to the vehicle is due for renewal and hence, these problems are avoided.

2.4 DATA RECORDING AND AVAILABILITY

Table 2-3 summarizes the test data recording methods and practices used in the various programs, plus the frequency with which the raw data would be available for analysis

The first column of Table 2-3 lists the method used to enter vehicle identification information (e g., VIN, license plate) into the test data

TABLE 2-3 DATA RECORDING AND AVAILABILITY

State	Vehicle ID/Test Results Entry Method[a]	Differentiate First Test from Retests	Tape Availability
AZ	Pre 7/86: Manual/On line 7/86+ On line/on line	No; every third test entered as first	Monthly, from state
CT	Manual/on line	No, every third test entered as first	Quarterly, from state
IL	On line/on line	Yes	Monthly, from state
KY	On line/on line	Yes	Quarterly, from contractor
MD	On line/on line	Yes; paid retests may be entered as first test	Annually, from state
WA	On line/on line	Yes	Semi-Annually, from state
WI	Bar code/on line	Yes, paid retests may be entered as first test	Monthly, from state

[a] Manual written or key punched by test operator, On line 1) Vehicle ID information from DMV records, 2) test results recorded by test apparatus.

bases. Vehicle identification is entered either manually or directly from a preexisting data base. The direct (on-line) data entry method is preferable as fewer errors are introduced and/or propagated, particular in recording the VIN. Direct data entry makes vehicle tracking, either during one test year or from year to year, more accurate. Only Connecticut relies on manual data entry while Arizona has converted to automated entry as of July 1986.

Two methods of recording tests and retests are used. One technique is to indicate as retests any test that is not the initial test for a vehicle in a given year This method is preferred for analysis as it simplifies the task of determining first test failure rates. The second technique is to record every paid test or every third test for a vehicle as a first test This method is not preferred due to the potential difficulty in recognizing first tests and calculating first test failure rates.

The third column in Table 2-3 lists tape availability frequencies These frequencies will be the minimum time between delivery of raw data tapes to EPA from the states involved, and will dictate the frequency with which EPA can perform analyses Two states, Maryland and Washington, do not produce master tapes on a schedule which lends itself to frequent data analysis.

Also listed in the third column are the sources of the raw test data All states will provide the raw data except Kentucky Kentucky's data can be provided from the contractor which administers the tests. In all cases, it appears that EPA's help will be required for the data to be released to a contractor. In addition, Connecticut requires a confidentiality of data agreement

2.5 SPECIAL FEATURES AND CONSIDERATIONS

Table 2-4 lists selected special features of each I/M program that can affect the usefulness of its data. The first column considers the final disposition of vehicles which failed all tests and retests during a given year. Knowledge of waivers, when granted, can allow year to year tracking of failed/waived vehicles. The second column lists the method used to handle retest records. Ideally, all test and retest data for a given vehicle would reside on one record with waivers indicated This method lends itself to the most accurate tracking of final vehicle results, which can be monitored year to year to determine emissions deterioration between inspections. Earlier analysis by EEA suggests that a population of vehicles fails at every inspection and is waived every year.

Only Arizona, Washington and Wisconsin merge test and retest records together. Of these three, only Arizona, and Wisconsin indicate the final status (pass/fail/waive) of each vehicle. Illinois, Kentucky and Maryland indicate waivers on the final test or retest record for each vehicle Connecticut and Washington keep separate files containing waiver information These files are not merged with the test results files so they are of no use to the study.

2.6 DATA FROM CALIFORNIA

Since California has separate standards and different engine families than the other 49-states, recognition of pattern case California families requires data from the state of California's inspection program. The data has two drawbacks - first, the program is decentralized and the quality of inspections unknown, second, the data does not contain the VIN number which is a basic requirement for identifying engine family However, there are some possible actions that one can take to enhance the value of the data

TABLE 2-4 SPECIAL FEATURES AND CONSIDERATIONS

State	Waiver Record Handling	Retest Record Handling
AZ	Waiver indicator	Merged with first test record
CT	Separate waiver file; merged with test file when available (still waiting for 1985 waivers)	Any retest records are separate entries
IL	Waiver indicator on final test record	Any retest records are separate entries
KY	Waiver indicator on final test record	Separate record by vehicle visit to inspection station
MD	Waiver indicator on final test record	Separate record by vehicle visit to inspection station
WA	Waiver not indicated on tape; physically tracked	Merged with first test record
WI	Waiver indicator on final test record	Results of up to 3 tests and/or retests per record

California requires that all vehicles up to 8,500 lb GVW in seven major metropolitan areas in California be inspected bi-annually, but this 8,500 lb GVW limit is based on registered GVW (as in many other states). The test used in an idle test, with 2500 rpm preconditioning. 1980 and later vehicles must meet standards for both the 2500 rpm test and the idle test. Loaded-mode cruise test standards are "on the books", but none of the regions require a loaded-mode test California is unique in having (normal) idle test standards that vary by technology type for 1980 and later cars, as shown below.

	HC	CO
No catalyst	150	2.5
Oxidation catalyst	150	2.5
Three-way open loop catalyst	150	1.2
Three-way closed loop catalyst	100	1.0

2500 rpm test standards are uniform at 220 ppm HC/1 2 percent CO for all technologies. It is not clear what percent of cars are misidentified and subjected to inspection at the wrong standards

Data entry on vehicle description is manual, and includes license plate, vehicle type, GVW, make abbreviation, model year, number of cylinders, engine size and odometer. EEA's examination of the records indicates far fewer than expected records qualified as an LDT As a result, it is possible that LDT's are being misclassified as far as the technology ? category. On average, slightly less than 10 percent of all records are classified as LDT's but registration records indicate that LDT penetration in California is over 25 percent for newer model years. All emission entries are automatic, with HC, CO, CO_2 and RPM recorded Test records are stored in a cassette tape at each mechanic station.

Data cassettes are collected on a monthly basis by a contractor and transcribed on to a mainframe computer by the Bureau of Automotive Repair (BAR) and its contracts. (This may be changed to process the records quarterly in the near future.) The total sample size is very large, with 600,000+ vehicles tested per month. In 1986, over 200,000 of these vehicles are 1980 and later models One advantage of the California data is that the Bureau of Automotive Repair already performs extensive cleaning of the data to eliminate records of calibration data, aborted tests, invalid tests, etc. Moreover, first test records and retest records for a given vehicle are merged and issuance of a waiver is noted. If vehicles have multiple first test records, these records are not merged, especially if they are from different stations. Yet another advantage of the California data is that the BAR utilizes SAS for doing its analysis and this will make the data easily adaptable to the processing system described in Section 4.

A major drawback of the California data base is the lack of the VIN number. We are currently using the VIN to determine:

- Make
- Model year
- Model name (carline)
- Engine displacement
- Aspiration,natural/turbocharged
- Gasoline/diesel
- GVW category (for trucks)

All but two of the variables are being manually recorded for each vehicle. The two variables not recorded are carline and aspiration. For over 80 percent of all vehicles, knowledge of the engine displacement, make and model year is sufficient to track engine family. (Of course, neither the VIN nor the above variables reveal California/Federal certification.) For the purpose of a general analysis that reveals most, but not all, pattern failures, this may be sufficient EEA does perceive a potential problem with not being able to distinguish between cars and light trucks. If the vehicle type information is poor, it is conceivable that unambiguous determination of emission control technology will be possible only in a small number of cases

The second method to overcome the data problem is by using the license plate information and Department of Motor Vehicles records to identify VIN. This would be expensive, as there are roughly 15 million light duty vehicles in the state, roughly a third of which are 1980 or later. BAR staff have contemplated this measure and believe that successful matches of correct VIN numbers will occur in about 70 percent of the cases. We cannot actually determine the quality of the VIN data unless a small sample of license plates are matched to VIN records and the VIN data examined This represents a possible area for additional exploration by EPA in the future

A very interesting feature about the California program is that it utilizes the 2500 rpm test for pass/fail determination and their idle cutpoints are more stringent than the EPA "207(b)" cutpoints. Either as a result of those factors, or due to other factors, California reports the highest failure rate for 1980+ cars in the nation. On average the failure rate (both tests combined) is approximately 25 percent, with even model year 1984 vehicles reporting failure rates of over 10 percent in 1985. These failure rates are much higher than those in other centralized I/M programs, even at the same cutpoints. In general, decentralized programs typically display low failure rates; California's failure rates are, therefore, surprising given that state officials have suggested that tampering and misfueling of vehicles is lower in California.

These factors suggest that California data could be useful to EPA, especially if matching license plate to VIN proves feasible.

2.7 SUMMARY

We have attempted to rank various aspects of each state's I/M program for its usefulness to the proposed analysis. This ranking is based on 10 categories; no weights have been placed on the categories but EPA may wish to weight the categories differently depending on immediate objectives. The categories are as below·

- Vehicles coverage to 8,500 lb - We have awarded 2 points for complete coverage, 1 for an intermediate point (such as 8,000 lb) and 0 for coverage up to 6,000 lb.
- Preconditioned idle test - We have awarded 2 points if all vehicles are subject to uniform preconditioning, 0 if there is no requirements, and 1 if it covers only failed vehicles. Uniform preconditioning is necessary to compare idle emissions and failure rates at cutpoints that differ from those in use
- Use of "207(b) standards - We have awarded 2 points for standard cutpoints, making failure rate comparisons simple, 1 point for a situation where 207(b) cutpoints are applied to only part of the 0-8,500 lb GVW fleet, and 0 for non-standard cutpoints
- Sample size - We have awarded 2 points if the total monthly sample is over 100,000 vehicles, 1 if it is between 50,000 and 100,000 and 0 if it is lower than 50,000.
- Data cleanliness - This refers to the presence of clean data for all fields. In general, manual entry of data on vehicle descriptions results in many errors, and is awarded 0 points. Fully automated systems that track vehicles through their registration records are awarded 2 points, while those are dependent on some manual inputs, e.g., test cycle number, are awarded 1 point
- Ability to distinguish first test - Given the fact that many vehicles have multiple records, either due to retests or several "first" tests, unambiguous determination of the first test in any calendar year is valuable. If the state can make this determination with accuracy, it simplifies processing requirements A score of 2 is provided if the test sequence variable is judged highly reliable, 1 if there is no on-line tracking of vehicles and 0 if there are known errors in this variable.
- Ability to gauge final outcome - This can be important if tracking waiver rates, or the repairability of pattern failures is an issue of interest. If all retests and the final outcome

(pass, fail, wavier, waiver type) of a given test sequence is available in the data, the score is 2 points' Availability of waiver data in a separate file that can be merged with the data is given a 1 point score, and data sets that do not show final outcome for failed vehicles are awarded 0 points.

- Data pre-sort - If the state pre-sorts all of the data to match records of each vehicle for test and retest, as well as for any unusual tests (e.g., change of ownership, multiple first tests) it is awarded 2 points, 1 point if only test and retest data is merged, and 0 if vehicle test records are unsequenced
- Retrieval time - This factor shows how long it takes to obtain the data after test completion. A score of 2 indicates data availability within 3 months, a score of 1 indicates data requiring 6 months and a score of 0 for a period longer than 6 months.

The scores for each of 7 centralized programs and California are shown in Table 2-5. Assuming all factors are weighted equally, Illinois, Kentucky and Wisconsin data appear to be the best, while Arizona, Connecticut, Maryland and Washington are less preferable. (Arizona's score is for the modernized system in use since July 1986.) Based on our previous experience, where we found problems with the Connecticut, Washington and (old) Arizona data to be of similar magnitude, the scores appear to reflect well their usefulness to the analysis.

TABLE 2-5

RANKING OF DATA USEFULNESS FOR SEVERAL I/M PROGRAMS

State	Total	Vehicle up to 8500 lb	Uniform preconditioning	207(b) Cutpoint	Sample Size	Date Cleanliness	Distinguish First Test?	Distinguish Final Outcome?	Data Pre-Sorted	Time Delay
AZ*	12	2	1	2	2	0	0	1	2	2
CT	10	2	1	2	2	0	0	1	0	2
IL	15	1	2	2	2	2	2	2	0	2
KY	14	2	2	2	0	2	2	2	0	2
MD	12	2	2	2	2	1	1	2	0	0
WA	11	2	2	0	1	1	2	0	2	1
WI	14	1	2	1	2	2	1	2	1	2
CA	12	2	2	0	2	0	1	2	1	2

* (revised program from July 1986)

3. DATA PROCESSING

3.1 OVERVIEW

The most resource intensive phase of the analysis is data processing. As described in Section 2, each state processes in the neighborhood of 100,000 to 200,000 records monthly. As a rule of thumb with newer model years, each model year accounts for 7 5 percent of all records. The LDV/LDT split is typically between 4·1 to 6.1 for a given model year. If we assume that EPA's interest in any calendar year is in the five model years covered by the emissions warranty, and the processing is on every six months of records, the data base of interest is somewhere between 225,000 to 450,000 records for each state. Analysis of such large data bases requires enormous amounts of computer time, and the slow turnaround of each run (usually overnight) makes it difficult to identify and correct errors.

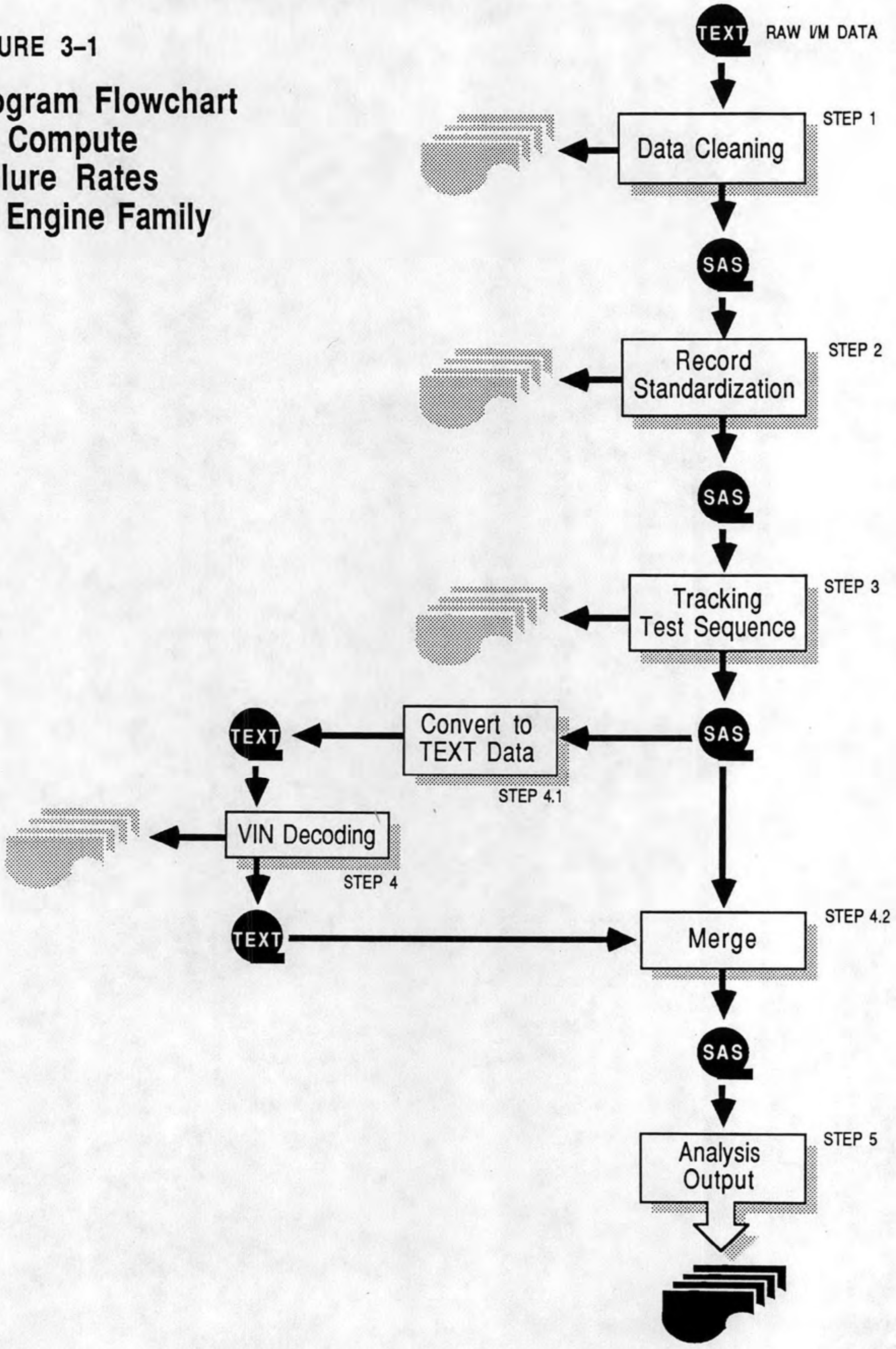
As a result, the data processing requirements are separated into a number of steps, with outputs at the end of each step to allow for error identification and correction. EEA's previous experience with I/M data programs suggests that these intermediate outputs are very important to the success of the project Accordingly the analysis of data has been divided into six steps:

- Data cleaning
- Standardization of variable/format
- Sorting and sequencing
- VIN decoding
- Merging all data on individual vehicles
- Analysis of failure rate and output

The steps are shown schematically in Figure 3-1 and discussed below.

FIGURE 3–1

Program Flowchart To Compute Failure Rates By Engine Family



3.2 STEP 1: DATA CLEANING

Any real world data tends to have errors; I/M data, especially in certain programs, have errors related to vehicle or emission variables, missing data fields or records generated from calibration and aborted tests that should not be used for analysis. The cleaning step removes such records by correcting or deleting them and also includes verification and conversion of data into an appropriate format.

The first activity is to verify the data by reading the raw tape and checking the values of vehicle descriptors (make, model year, type, cylinders, odometer) and emission readings to make sure they do not exceed the allowable range or contain blank fields Another item included in the verification is the test result variable ("P" or "F"); this can be computed by comparing the emission readings to standards associated with the particular model year (and, in some cases, the number of cylinders). This step assumes that the tape copy of data received from an I/M program is in the format agreed upon and the variables are properly understood.

The second activity is the actual process of removing data records that are incorrect, have missing fields or are not relevant for this analysis In this step, tests of only light-duty vehicles (LDV) and light-duty trucks (LDT) are retained, and an additional step of removing all diesel LDV/LDT may be performed if there is an appropriate indicator field In particular, the vehicle type indicator (absent in some states) is retained for error checking after VIN decoding of vehicle type In addition, aborted tests, calibration test and incomplete test records are deleted.

The third activity involves assigning a separate computed pass/fail variable (distinct from the one recorded) for results checking at the end of other processing steps This is very useful in determining if other factors are being used in the program to pass or fail vehicles (e.g., a tolerance on the standard that would pass vehicles very slightly above standards or an underhood inspection that fails vehicles passing the emission test).

Statistics on the total number of records and its breakdown by model year, vehicle type, test month/year, make and test results before and after processing are recommended outputs for this step. The statistics are useful in determining what percentage of data is being rejected due to cleaning, and any bias in record rejection (i.e., failed vehicles having a larger percentage of records rejected than passed vehicles) A very high record rejection rate or large bias in record rejection may require acquisition of a new data tape or discussions with the program managers to pinpoint the causes of the errors

Although this processing step appears routine, it has been EEA's experience that this step is necessary but tedious. In addition, this cleaning step cannot be "standardized" as the types of errors and requirements for record rejection vary from program to program. As an example, the coding of emission values can be in percent CO or hundredths of a percent of CO, and format specifications may neglect to mention the units Moreover, EEA has had the experience where there are unannounced format changes in the program leading to considerable confusion Therefore, this step cannot be automated but instead requires considerable intervention on the part of both a programmer and an analyst

3.3 STEP 2: RECORD STANDARDIZATION

It is anticipated that data from several different I/M programs will be processed through the VIN decoder and analyzed Accordingly, this step deals with:

- Dropping unnecessary variables
- Developing a standard format for variables of interest

- Standardizing alpha-numeric variables
- Reading into SAS

Several variables recorded in the data set at any I/M program will be irrelevant to the analysis required in this assignment. They include records of inspection sticker number, tax codes, safety test results, repair cost information, etc. Moreover, since we plan to analyze only 1981 and later vehicles (or the last five model years), we can delete all unnecessary records and fields to minimize data storage requirements and processing costs.

The second activity in this step involves creating a standard format for all variables of interest. While this process is relatively straightforward, one area of particular concern is the test procedure and the several variables in the procedure and in pass/fail requirements. The number of HC/CO emission records vary according to the procedures in-use which include·

- All vehicles subjected to preconditioning, only one test at idle
- Unpreconditioned idle test, with preconditioning and a second idle test only for failed vehicles
- Unpreconditioned idle, high idle/loaded mode and second idle tests for all vehicles for a total of three tests.

Pass/fail determination can be based upon any one, two or all three test modes in some states; additionally, test and pass/fail requirements can vary by model year. There is variation in the types of preconditioning and states may change the test over time These test specific emission records must be carefully tracked and the format must allow specification of any combination of test mode and pass/fail criteria. In previous work efforts for EPA, EEA suggested a standard format, but we now believe this should be enhanced by an additional variable that provides information on which test results are used to determine pass/fail, and to distinguish between blank, missing and "zero" fields

Standard format specification and care in conversion of alpha-numeric variables apply primarily to the VIN, license plate, and MAKE codes. Problems can arise in reading such variables and field length specifications are critical to avoid truncation errors. The MAKE code is required primarily for error checks in sorting, as detailed below. Typically, no standard abbreviations are used for makes and multiple alternatives are used in the same state for designating the same make. Moreover, same abbreviations lead to confusion - a common one is the use of "MERC" to denote both Mercury and Mercedes-Benz In the past, EEA has utilized a dictionary that maps up to 99 percent of non-standard abbreviations into standard abbreviations as MAKE codes. This dictionary is constructed by printing out all variables in MAKE in the raw data tape, and assigning non-standard formats to standard codes This time consuming effort may not be necessary if the VIN data is clean, as described in the following subsection.

The step 2 processing reports will generate statistics on a number of records and statistics on each field for blanks or missing data. If required, a MAKE code frequency table and a report on makes not mapped into standard code can also be provided.

3.4 STEP 3: TRACKING TEST SEQUENCES

This step is required to separate first test and retest records as well as to track multiple "first" tests. As described in Section 2 of this report, most centralized I/M programs have a variable to indicate first test or retest, but EEA's experience has been that the variable is not completely reliable. For example, in Arizona and Connecticut, every third test (second retest) is counted as a first test

The data base must first be sorted to match all available records as a single vehicle. Two types of test must be distinguished

- Multiple first tests
- Multiple retests.

Multiple first test records can occur in an I/M program if motorists go to separate I/M stations on the same day after failing a first test and decide the vehicle can pass on a second try at a different station. I/M locations like Kentucky have on-line computers that will prevent motorists from claiming a second "first test", but many I/M programs cannot recognize such vehicles as having already completed a first test Multiple first test records can also occur over the course of six months or a year if vehicles are required to go through both an annual inspection and an inspection at change-of-ownership.

Retest records are easily confused with first test records as many owners let the allowable repair period elapse before they appear for their retest. In states with change of ownership inspections, it is sometimes difficult to exactly distinguish which tests are retests. In addition, vehicles with multiple retests have records that are more susceptible to incorrect data entry (especially in I/M programs with manual data entry) Tracking of first test and retest is important for two reasons·

- Records confusion exists <u>only</u> for vehicles failing the first test, and their elimination will result in biased calculations.
- In the interest of 207(b) warranty enforcement, EPA may need to know the final outcome of test sequences

Sorting of all records by data for each vehicle is required for assigning test sequence number. For the purposes of this work effort, we have <u>not</u> pursued the algorithm for assigning the correct retest or multiple first test number, and instead focused this effort into simply determining with as much accuracy as possible, the first test for a given vehicle in a given year

Sorting can be based solely on VIN, in states with manual data entry, VIN keypunch errors may result in a poor match of records EEA has used VIN <u>or</u> (MAKE <u>and</u> MYR <u>and</u> LICENSE PLATE) as a second sorting criterion All three variables must be matched together because license plates need not be unique between commercial and non-commercial vehicles and, in some states, the plates can be transferred from one vehicle to another Contrasting the number of record matches using the two methods is a useful check of VIN keypunch errors. This, of course, requires extra effort in Step 2 for MAKE codes standardization

As an example, we utilized a sample of records from Connecticut, illustrating the range of variation observed. The sorting performed was in two different ways: first by VIN only and second by license plate and MAKE/MODEL YEAR. Table 3-1 illustrates the results in the matrix of record counts by the two methods. If VIN sorting produces a record count for any particular vehicle of N, and sorting by license plate/ MAKE/MYR a record count for the same vehicle of M, ideally M should equal N. However, a small percentage of cars sorted by the second method show values of M lower than N, but in no case does M exceed N This indicates that sorting by VIN is superior in all instances to sorting by license plate/MAKE/MYR if the data is from a limited time period. Over longer period such as two years, this may not be true.

Kentucky does not record license plate but should theoretically have a test number variable that is very reliable Table 3-2 shows the results of the VIN sort number N, as a function of Kentucky's test number -- 1, 2, and S (greater than 2) Clearly, for N=1, it is possible to have vehicles with a higher Kentucky test number if their previous records are in an earlier data tape Surprisingly, 12 percent of test records for N=2 was labelled by Kentucky as a first test, indicating potential deficiencies in the system. The table illustrates the need for the sorting step even when we analyze data from a highly computerized I/M/ program

TABLE 3-1

CONNECTICUT 1984 QUARTER 1 AND QUARTER 2 VIN AND PLATE COUNTS
N = TEST SEQUENCE BY VIN, M = TEST SEQUENCE BY PLATE

TABLE OF N BY M

```
N         M

FREQUENCY|
 PERCENT |
 ROW PCT |
 COL PCT |       1|       2|       3|       4|       5|       6|       7|       8|       9|      11|  TOTAL
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       1 | 143683 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |      0 | 143683
         |  78.87 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  78.87
         | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         |  99.73 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       2 |    233 |  35648 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |  35881
         |   0.13 |  19.57 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  19.70
         |   0.65 |  99.35 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         |   0.16 |  99.59 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       3 |     80 |     66 |   1196 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |   1342
         |   0.04 |   0.04 |   0.66 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.74
         |   5.96 |   4.92 |  89.12 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         |   0.06 |   0.18 |  96.69 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       4 |      8 |     42 |      7 |    435 |      0 |      0 |      0 |      0 |      0 |      0 |    492
         |   0.00 |   0.02 |   0.00 |   0.24 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.27
         |   1.63 |   8.54 |   1.42 |  88.41 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         |   0.01 |   0.12 |   0.57 |  93.95 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       5 |     12 |      3 |      9 |      2 |     88 |      0 |      0 |      0 |      0 |      0 |    114
         |   0.01 |   0.00 |   0.00 |   0.00 |   0.05 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.06
         |  10.53 |   2.63 |   7.89 |   1.75 |  77.19 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         |   0.01 |   0.01 |   0.73 |   0.43 |  83.02 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
TOTAL      144073    35796     1237      463      106       65       35       30       24      349   182179
            79.08    19.65     0.68     0.25     0.06     0.04     0.02     0.02     0.01     0.19   100.00
```

TABLE 3-2

KENTUCKY DATA SORTING

EEA TEST NUMBERS VERSUS KENTUCKY TEST NUMBERS

```
           TABLE OF N BY TST_SEQ

N           TST_SEQ

FREQUENCY|
 PERCENT |
 ROW PCT |
 COL PCT |S       |1       |2       |  TOTAL
---------+--------+--------+--------+
       1 |     24 |  33442 |    408 |  33874
         |   0.07 |  93.45 |   1.14 |  94.66
         |   0.07 |  98.72 |   1.20 |
         |  14.72 |  99.36 |  20.76 |
---------+--------+--------+--------+
       2 |     39 |    209 |   1530 |   1778
         |   0.11 |   0.58 |   4.28 |   4.97
         |   2.19 |  11.75 |  86.05 |
         |  23.93 |   0.62 |  77.86 |
---------+--------+--------+--------+
       3 |     94 |      7 |     27 |    128
         |   0.26 |   0.02 |   0.08 |   0.36
         |  73.44 |   5.47 |  21.09 |
         |  57.67 |   0.02 |   1.37 |
---------+--------+--------+--------+
       4 |      5 |      0 |      0 |      5
         |   0.01 |   0.00 |   0.00 |   0.01
         | 100.00 |   0.00 |   0.00 |
         |   3.07 |   0.00 |   0.00 |
---------+--------+--------+--------+
       5 |      1 |      0 |      0 |      1
         |   0.00 |   0.00 |   0.00 |   0.00
         | 100.00 |   0.00 |   0.00 |
         |   0.61 |   0.00 |   0.00 |
---------+--------+--------+--------+
TOTAL         163    33658     1965    35786
             0.46    94.05     5.49   100.00
```

3.5 STEP 4: VIN DECODING

Decoding the VIN to obtain engine family designation and emission control system type (fuel control, catalyst, or secondary air fuel system type) has been developed by EEA and is now available as a stand alone program (For a description, see "VIN Decoder· User's Guide" EEA Report to the EPA, September 1986.) The current VIN decoder is capable of analyzing and decoding VIN for model years 1981-1984 light duty vehicles and light-duty trucks. Another product of the VIN decoder is an error report that allows tracking of the number of records with VIN errors.

For this step, EEA recommends that only the license plate, make and model year be retained with VIN in a separate data tape in TEXT format for input to the VIN decoder thus minimizing memory requirements and input/output processing. The operation of the VIN decoder as a unit is straightforward, and the error analysis is output as required. A sample of Kentucky and Connecticut calendar year 1984 data was processed to reveal the typical percentages of record retention. Table 3-3 shows the VIN numbers decoded for vehicles designated as MYR 1981-1984 in Connecticut. As can be seen, only 82.8 percent of VIN's are successfully decoded. Two major error types - 08 and 11 - account for most of the VIN errors. Error code 08 arises from failure of the validity test, and 11 arises from non-standard VIN format, potentially as a result of truncation of the VIN

Table 3-4 shows the results of VIN decoding for Kentucky data - 94.85 percent are successfully decoded, and the major error type (code 02) arises from the particular engine key not being found in the table One explanation for this is that running changes are not being incorporated into the VIN decoder's certification data tape at the current time. The examples show that the VIN decoding success rate is likely to vary from

```
                              VIN DECODING OF CT1984

                              TABLE OF MYR BY ERRLVL

MYR        ERRLVL

FREQUENCY|
 PERCENT |
 ROW PCT |
 COL PCT |00       |03       |05       |07       |08       |09       |10       |11       |12       |  TOTAL
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
       0 |       0 |       0 |       0 |       0 |       0 |       0 |       0 |   20082 |     324 |  20406
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |   10.89 |    0.18 |  11.07
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |   98.41 |    1.59 |
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |  100.00 |  100.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      80 |      28 |       1 |       0 |       0 |       0 |       0 |       0 |       0 |       0 |     29
         |    0.02 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |   0.02
         |   96.55 |    3.45 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |
         |    0.02 |    0.38 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      81 |   45428 |      44 |      31 |     146 |    2654 |      88 |     528 |       0 |       0 |  48919
         |   24.64 |    0.02 |    0.02 |    0.08 |    1.44 |    0.05 |    0.29 |    0.00 |    0.00 |  26.54
         |   92.86 |    0.09 |    0.06 |    0.30 |    5.43 |    0.18 |    1.08 |    0.00 |    0.00 |
         |   29.76 |   16.54 |   13.66 |   28.68 |   35.34 |   20.47 |   22.33 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      82 |   48165 |      15 |      40 |     198 |    2492 |     109 |     502 |       0 |       0 |  51521
         |   26.13 |    0.01 |    0.02 |    0.11 |    1.35 |    0.06 |    0.27 |    0.00 |    0.00 |  27.95
         |   93.49 |    0.03 |    0.08 |    0.38 |    4.84 |    0.21 |    0.97 |    0.00 |    0.00 |
         |   31.56 |    5.64 |   17.62 |   38.90 |   33.19 |   25.35 |   21.23 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      83 |   55735 |     152 |     153 |     150 |    2145 |     197 |    1146 |       0 |       0 |  59678
         |   30.24 |    0.08 |    0.08 |    0.08 |    1.16 |    0.11 |    0.62 |    0.00 |    0.00 |  32.37
         |   93.39 |    0.25 |    0.26 |    0.25 |    3.59 |    0.33 |    1.92 |    0.00 |    0.00 |
         |   36.52 |   57.14 |   67.40 |   29.47 |   28.57 |   45.81 |   48.46 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      84 |    3270 |      54 |       3 |      15 |     218 |       7 |     189 |       0 |       0 |   3756
         |    1.77 |    0.03 |    0.00 |    0.01 |    0.12 |    0.00 |    0.10 |    0.00 |    0.00 |   2.04
         |   87.06 |    1.44 |    0.08 |    0.40 |    5.80 |    0.19 |    5.03 |    0.00 |    0.00 |
         |    2.14 |   20.30 |    1.32 |    2.95 |    2.90 |    1.63 |    7.99 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
      85 |       0 |       0 |       0 |       0 |       0 |      29 |       0 |       0 |       0 |     29
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.02 |    0.00 |    0.00 |    0.00 |   0.02
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |  100.00 |    0.00 |    0.00 |    0.00 |
         |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |    6.74 |    0.00 |    0.00 |    0.00 |
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
TOTAL       152626       266       227       509      7509       430      2365     20082       324   184338
             82.80      0.14      0.12      0.28      4.07      0.23      1.28     10.89      0.18   100.00
```

TABLE 3-3

TABLE 3-4

```
                                   VIN DECODING OF KY1984                   15:56 FRIDAY, SEPTEMBER 26,

                                   TABLE OF MYR BY ERRLVL

MYR       ERRLVL

FREQUENCY|
 PERCENT |
 ROW PCT |
 COL PCT |00      |03      |05      |07      |08      |09      |10      |11      |12      |  TOTAL
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       0 |      0 |      0 |      0 |      0 |      0 |      0 |      0 |    162 |      5 |    167
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.56 |   0.02 |   0.57
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  97.01 |   2.99 |
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 | 100.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
      81 |   7752 |    290 |      0 |      7 |     66 |      7 |     15 |      0 |      0 |   8137
         |  26.62 |   1.00 |   0.00 |   0.02 |   0.23 |   0.02 |   0.05 |   0.00 |   0.00 |  27.94
         |  95.27 |   3.56 |   0.00 |   0.09 |   0.81 |   0.09 |   0.18 |   0.00 |   0.00 |
         |  28.06 |  42.96 |   0.00 |  29.17 |  36.87 |  12.28 |   5.14 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
      82 |   7733 |    195 |     35 |      7 |     47 |     19 |      8 |      0 |      0 |   8044
         |  26.55 |   0.67 |   0.12 |   0.02 |   0.16 |   0.07 |   0.03 |   0.00 |   0.00 |  27.62
         |  96.13 |   2.42 |   0.44 |   0.09 |   0.58 |   0.24 |   0.10 |   0.00 |   0.00 |
         |  27.99 |  28.89 |  33.33 |  29.17 |  26.26 |  33.33 |   2.74 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
      83 |   8478 |     81 |     57 |      8 |     48 |     21 |    117 |      0 |      0 |   8810
         |  29.11 |   0.28 |   0.20 |   0.03 |   0.16 |   0.07 |   0.40 |   0.00 |   0.00 |  30.25
         |  96.23 |   0.92 |   0.65 |   0.09 |   0.54 |   0.24 |   1.33 |   0.00 |   0.00 |
         |  30.69 |  12.00 |  54.29 |  33.33 |  26.82 |  36.84 |  40.07 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
      84 |   3664 |    109 |     13 |      2 |     18 |      6 |    152 |      0 |      0 |   3964
         |  12.58 |   0.37 |   0.04 |   0.01 |   0.06 |   0.02 |   0.52 |   0.00 |   0.00 |  13.61
         |  92.43 |   2.75 |   0.33 |   0.05 |   0.45 |   0.15 |   3.83 |   0.00 |   0.00 |
         |  13.26 |  16.15 |  12.38 |   8.33 |  10.06 |  10.53 |  52.05 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
      85 |      0 |      0 |      0 |      0 |      0 |      4 |      0 |      0 |      0 |      4
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.01 |   0.00 |   0.00 |   0.00 |   0.01
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |
         |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   7.02 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
TOTAL       27627      675      105       24      179       57      292      162        5    29126
            94.85     2.32     0.36     0.08     0.61     0.20     1.00     0.56     0.02   100.00
```

80 to 95 percent. At 80 percent, some formal steps are required to check for the high error rate. If the VIN errors are biased towards failed vehicles, then deletion of record with VIN error could substantially alter failure rate computations. One check method for data with high error rates would be to tabulate the VIN error for "passed" and "failed" vehicles separately and check the statistics for bias.

Two other administrative problems have been noted by EEA. The certification data tape for any specific model year released by EPA (usually in March) is based on pre-model year data, and does not contain any running changes made by the manufacturer during the model year. Apparently, EPA has a separate file in which running changes for all major manufacturers except GM are maintained. The file cannot be easily merged with certification data as the formats are not similar. At this point, the resolution of the running change problem does not appear simple, and may not be able to be resolved

The second administrative problem relates to vehicles whose title specific model year, VIN decoded model year and engine family model year are inconsistent EEA was not able to resolve why this problem exists, but has learned informally that there may be some confusion at the close of one model year and the beginning of the next, between VIN model year and engine family model year. In general, this has resulted in only a small number of vehicles (less than 1 percent of the sample) being potentially misclassified for engine family designations.

EPA is aware that VIN decoding does not allow recognition of 49-State versus California certification Decoding by the manufacturers has allowed EEA to establish that even in a neighboring state to California (like Arizona), the number of California vehicles is less than 4 percent In other states further away from California, it is anticipated that California vehicles are less than 1 percent of the population. Moreover, 49-state and California cars have, in recent years, become nearly technologically identical and differ only in calibration. As a result, we believe this issue is not of significant concern except in the case of California vehicles where 49-state vehicles are estimated to be 10-15 percent of the vehicle population.

Finally, EPA has been interested in failure rate by engine family and transmission type, as certain models with automatic transmissions have been reported as pattern case failures. We examined the VIN code and determined that only four manufacturers - AMC, Honda, Renault and Subaru - entered transmission type information in the VIN. As a result, computation of failure rates at this level of detail is not possible.

3.6 STEP 5: ANALYSIS OUTPUT

After merging the VIN decoder outputs with the emissions test data, the generation of failure rates by engine family is a straightforward step, requiring only the cutpoints and the test results (T1, T2, T3) combinations to be considered for determination of failure. One advantage of utilizing SAS is that it can provide failure rate statistics by engine family and by other levels of aggregation such as emission control type, manufacturer, vehicle type, with very little additional programming The generation of output tables in SAS is less convenient, but EEA already has extensive programs to generate tables of failure rates at different strata

A second advantage in utilizing SAS is that the output data file can be directly tapped into for further statistical analysis to determine which engine families are pattern case failures The methods, and their availability in SAS are addressed in Section 4 of this report.

4. STATISTICAL ANALYSIS FOR IDENTIFYING PATTERN CASE FAILURES

4.1 INTRODUCTION

EPA is interested in identifying engine families that may be failing at rates significantly higher than average on the state vehicle inspection/maintenance test. An engine family corresponds to a unique make/model/engine size/emission technology, and is used by EPA to determine certification to standards and for recall. A failure is recognized when tail pipe HC and CO idle emission concentrations exceed a given set of cutpoints. EPA is interested in failure rates computed for at least two sets of cutpoints -- 100 ppm HC/0.5% CO and 220 ppm HC/1.2% CO.

Once the failure rates by engine family at each cutpoint are computed based on each individual state's I/M data, there are some additional complications in comparisons between states. Each state has a slightly different I/M test procedure that can give rise to differences in failure rates. In addition, climatic variables, such as temperature, can also influence failure rates. Given all of these effects, the question is what statistical test or tests should be employed to recognize high failure rate families? How much data is required to recognize these families?

Given this situation, there are five topical questions of interest to the EPA. They are:

1. Defining "high failure rate." EPA has used a χ^2 test comparing each family's failure rate to the fleet average failure rate. Is this appropriate if the high failure rate families are a significant portion of the fleet, thus biasing the average?

2. Since there are different technologies used to meet standards -- e.g., carburetor versus fuel injection -- should the failure rate for each family be compared to others in the same technology group?

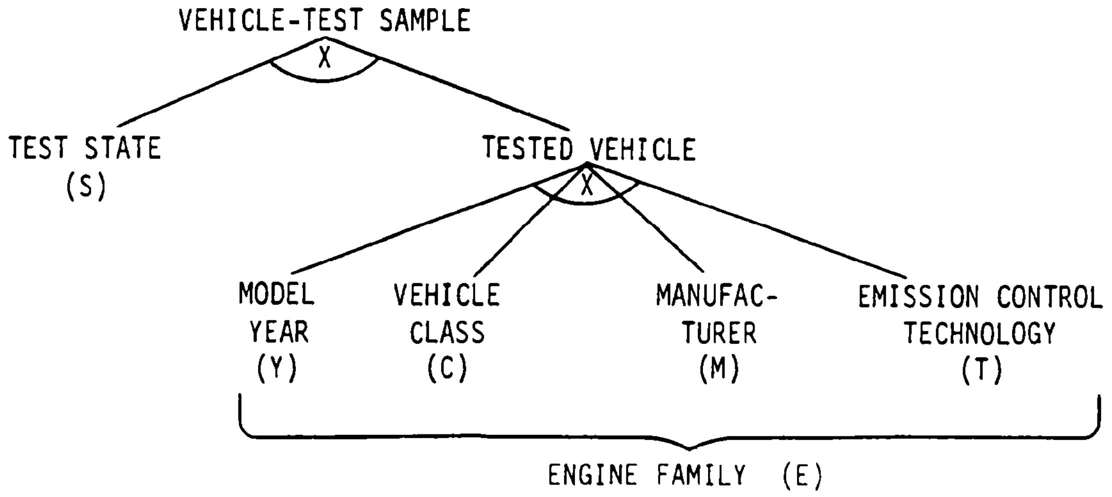
3. For many of the newer model years, the fleet average failure rates are very low -- 1 to 2 percent. How should test methods and sample sizes be structured in the comparisons?

4. What is the most appropriate statistical test to compare a given engine family's failure rate across state specific data? How can data from different states be combined to increase resolution?

5. Should the recommended statistical tests be performed separately for each set of cutpoints?

4.2 DESCRIPTION OF THE DATA AND UNDERLYING ASSUMPTIONS

For the purposes of this inquiry, the available processed data can be described as a collection of distinct vehicle-test samples, each sample characterized by a sample size (number of vehicles tested) and two test results: (1) the number or percent of vehicles failing the test under criterion 207(b), i.e., cutpoints 220 ppm HC/1.2% CO; and (2) the number or percent failing under cutpoints 100 ppm HC/0.5% CO. The qualities that define a distinct vehicle-test sample are: the unique engine family (as certified by EPA) to which the vehicles belong and the state in which the tests occurred. Engine families are, further, classifiable by model year, vehicle class (light-duty vehicle, LDV or light-duty truck, LDT), manufacturer, and emission control technology type. The block structure for the vehicle-test samples may thus be diagrammed as shown below:



The structure may also be expressed in a conventional algebraic notation by Sx((YxCxMxT) → E). The x symbol denotes crossing of "treatments" while the → symbol denotes nesting. Thus, within each (Y, C, M, T) combination there will be zero or more distinct engine families (E), but the individual engine families within one (Y, C, M, T) combination bear no relation to engine families within any other combination.*

The attached page from an EPA report illustrates the nature of the data for state (S) = Washington (Seattle), model year (Y) = 1982, vehicle class (C) = light duty vehicle (LDV), four particular manufacturers (M) = Nissan, etc. and the individual 1982 LDV engine families (F) of these manufacturers. Each engine family belongs to a specific emission control technology type (T), and these are written in for most of the families on the page. Each line thus represents a specific vehicle-test sample and includes the three essential numerical outputs for the inquiry: the number in the sample (N) and the calculated failure rates P_1 and P_2 corresponding to the two designated sets of cutpoints, 220 ppm HC/1.2% CO and 100 ppm HC/0.5% CO.

A few preliminary observations. Sample size N varies over a tremendous range. Although engine family entries with $N < 10$ are typically the result of erroneous decoding of the vehicle identification number (VIN) and can usually be ignored, the differences in precision of the estimated failure rates are still very great. In fact, some popular domestic manufacturer engine families have sample sizes exceeding 10,000 in some state programs. For example, 13 distinct emission control technology types (T) have been defined in the EPA report[1] for characterizing all model year 1982 vehicles, but no (C, M) combination contains engine families falling into more than four technology types. Furthermore, the number of engine families corresponding to a particular (Y, C, M, T) combination that is represented also varies. Thus, the block structure is quite sparse and unbalanced for several reasons -- the restriction to a small subset of all possible (Y, C, M, T) combinations, variable numbers of engine families per represented combination, and widely varying sample sizes among the individual families.

* There is an exception to this statement. Some engine family certifications are carried over from one model year to the next. The possibility of identifying some engine families across model years therefore exists, but will be ignored in the present analysis.

TABLE 4-1

: FAILURE RATE SUMMARY BY ENGINE FAMILY

MODEL YEAR 1981 LIGHT-DUTY VEHICLES ---------- INITIAL TESTS ----------

		N	ARIZONA FAILURE RATE(%)	207(B) FAILURE RATE(%)	100/0.5 FAILURE RATE(%)	Technology Type
				P1	P2	
AMERICAN MOTORS						
BAM151V2BC4	BAM151V2FC1	362	1.9	1.9	3.6	CARB/OXD/PMP
BAM258V2HP7	BAM258V2HP7	2020	9.3	11.6	18.5	CARB/3CL/PMP
CHRYSLER CORP.						
BCR1.7V2HJ1	BCR1.7V2HJ1	823	6.7	7.8	12.9	CARB/3CL/OXD/PMP
BCR2.2V2HA5	BCR2 2V2HU8	3508	4.2	5.6	8.7	
BCR2.6V2BJ2	BCR2.6V2BL4	1940	1.5	2.7	5.5	CARB/OXD/PLS
BCR3.7V1BA0	BCR3.7V1HE5	793	3.8	5.2	8.7	CARB/OXD/PMP
BCR5.2V2HJ4		774	2.1	3.7	12.3	CARB/3CL/OXD/PMP
BCR5.2V4HC1	BCR5.2V4HC1	116	12.9	15.5	28 4	
BCR5.2V9FAX	BCR5.2V9FF6	155	0.6	1.9	3.2	FI/3CL/OXD/PMP
FORD MOTOR CO.						
1.6AP	1.6APC	6108	7.2	9.4	19.5	CARB/3 WAY/OXD/PMP
2.3AHF		799	5.9	6.3	8.5	
2.3AX	2.3AX	2132	7.4	8 3	11.1	CARB/3CL/OXD/PMP
3.3GQF	3.3GQF	10621	6.9	8 9	16.0	CARB/3 WAY/OXD/PMP
4.2/5.0AAC		1111	2.3	3.1	6.2	CARB/3CL/OXD/PMP
4.2/5.0GCC/ACC		29	24 1	27 6	34.5	
4.2/5.0GCC/GCF	4.2/5.0GCC/ACC	142	9.9	10.6	16.9	
4.2/5.0GCF	4 2/5.0AAC	2248	13.5	16 1	23.5	
4.2/5.0MAF	4.2/5 0GCC/ACC	1743	3 7	4.8	9.0	CARB/3WAY/OXD/PMP
5.0CCF	5 0CCC	1274	2 0	6 0	22 9	CARB/3CL/OXD/PMP
5.8WBPF	5.8WAXC	407	9 8	13.3	19.7	
GENERAL MOTORS						
11C2NDM/NN	11C2NDM/NN	9595	2.4	2.9	4.7	
11D2AC		3153	2.1	3.4	11.1	
11E2AC		7839	2 8	3.6	11.2	All CARB/3CL/OXD/PMP except:
11L4AC		4295	3 8	5.2	14.9	
11L4ACJ		281	3 6	4.6	14.2	
11W2TNQZ	11W2TNQZ	7574	11.6	15.4	32.5	CARB/3CL/PLS
12H2AD	12H2AD	1123	1.8	2.4	6.4	
12S4AB	12S4AB	113	1.8	3.5	12.4	
12S4ABD	12S4ABD	154	4.5	5.8	10.4	
12X2NN	12X2NN	8068	0.8	1.2	2.6	
13H2AE	13H2AEJ	2295	1.7	2.7	8.6	
13Y4AR		3904	2.5	3.6	8.4	
14E2TM	14E2TM	14517	6.4	8.8	30 6	CARB/3CL/PMP
14E4NBD		204	4.4	6.4	23.5	
14F4AE	14F4AEJ	1322	1.9	3.0	7.5	
16T5ADB	16T5ADBJ	1587	0.5	0.8	3.6	FI/3CL/OXD/PMP
16T5ARB/DB		1745	0.7	1.7	3.8	

Other variables may be recorded which characterize individual tested vehicles and which could very well be correlated with measured emission levels. Notably among these are: odometer mileage, age (calendar year - model year), and month of test. For purposes of the present inquiry, however, these factors will be ignored and it will be assumed that each vehicle-test sample, i.e., (S, Y, C, M, T, E) combination is a sample from a homogeneous population. The population is then fully described statistically by two parameters: p_1 and p_2, the probabilities of failing cutpoint sets 220/1.2% and 100/0.5%, respectively. The response data P_1 and P_2 represent estimates of these underlying parameters. There is a slightly more unifying way of viewing the two responses which derives from the fact that the 220/1.2% failures are a subset of the 100/0.5% failures, i.e., one criterion subsumes the other. This is the categorical response viewpoint which says that the result of a test puts the vehicle into one of three mutually exclusive categories: pass 100/0.5%, fail 100/0.5% but pass 220/1.2%, and fail 220/1.2%. The associated probabilities are $1 - P_2$, $P_2 - P_1$, and P_1.

We proceed, next, to consider the five questions posed in the Statement of Work. The focus in Question 1 ("high failure rate") is on comparing engine families within a "fleet," without specific reference to explanatory factors. The details of the block structure defined above will not be involved. It is in response to Question 2, which raises the issue of technology type influence, where we will introduce an approach for assessing the significance of effects attributable to various factors represented in the block structure. Our discussions in Question 3 will consider the interplay of sample size, diminishing failure rates, and correspondingly lowered criterion for "high failure rate" in affecting the power with which high failure rate families can be successfully identified. In dealing with Question 4 on across-state comparisons, we will expand on the approach suggested in Question 2 which should also result in attaining increased explanatory power. Suggestions for handling multiple-valued response, as requested in Question 5, will be made.

4.3 QUESTION NO. 1: DEFINING "HIGH FAILURE RATE"

Consider a data set of k vehicle-test samples (engine families) denoted by $(n_1, P_1), \ldots, (n_k, P_k)$ where n_i is the i^{th} sample size and P_i is the failure rate within the i^{th} sample calculated with respect to a single set of cutpoints. The issue of multiple sets of cutpoints is reserved for Question 5. It is presumed that this data set is restricted to a particular state, a particular vehicle class (LDV or LDT), and a particular model year. Even though many different manufacturers are involved and they apply a variety of emission control technologies, in principle, one might have expected a fairly homogeneous collection of true failure rates because the test method, the distribution of environmental conditions, statutory emission standards for vehicle certification, and the state-of-the-art apply uniformly over this set of engine families. In practice, one finds a considerable spread of estimated rates. The problem posed is to quantify the notion of "high failure rate" and to describe a procedure for identifying the subset of engine families which can confidently be said to have high failure rates.

A concrete example provides a useful framework for discussion. In the accompanying figure are plotted (in rank order) the estimated 220/1.2% failure rates ±1 standard error for 20 engine families more or less serially selected from the first three listed manufacturers in the failure rate summary table for model year 1982 LDV's in Arizona. No one is likely to argue about calling families 18-20 high failure rate families. What about families 16 and 17? Their estimates are distinctively high, but, because of small sample size, comparison tests with any of the smaller-rate families are not likely to show any statistically significant difference. What about families 1-15? There is no intuitively obvious way of partitioning that group into "normal" and "high" rate subgroups; still, statistical comparison tests would likely show 11-14 significantly different from 1 and 2.

What are some of the classical statistical methods for multiple comparisons among engine family "treatments" which could be applied here? Many methods are inapplicable because of unequal sample sizes. One in particular is Duncan's multiple range test, even with Kramer's extension to unequal sample

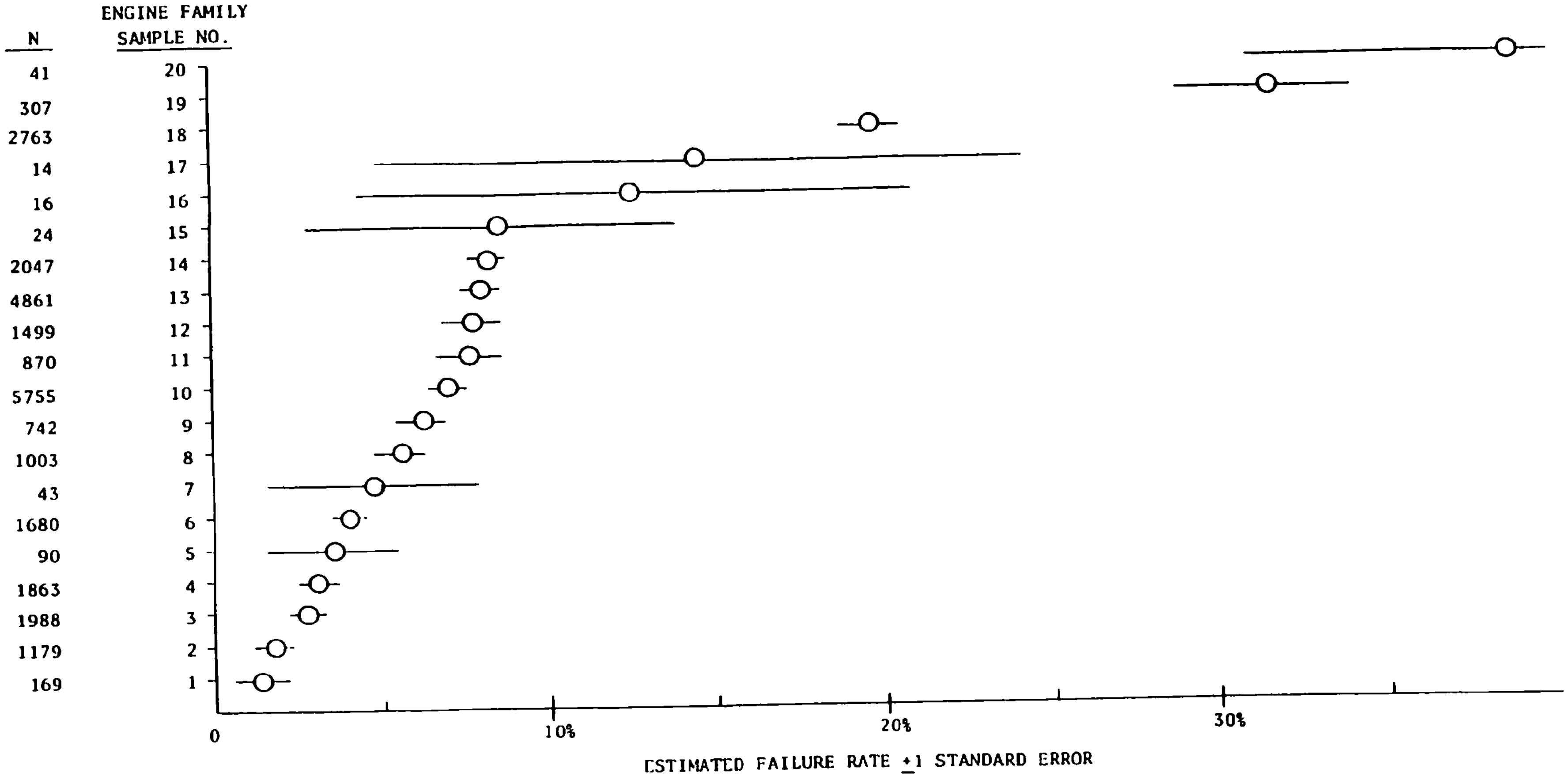


FIGURE 4-1 EXAMPLE DISTRIBUTION OF FAILURE RATES

sizes.[1] The problem with the extension is that it can't properly handle large variations in sample size which is what we need here. This is a shame because the SAS statistical software package[2] has Duncan's procedure with Kramer's extension. A very well-known method due to Scheffe and the "multiple-t" method are both applicable.[3] For both methods, it is necessary to compute the within-treatments mean square, MS_w, which can be expressed as a pooled variance, namely,

$$MS_w \quad \frac{\sum_{i=1}^{k} (n_i - 1)\, s_i^2}{\sum_{i=1}^{k} (n_i - 1)} = \frac{1}{n - k} \sum_{i=1}^{k} (n_i - 1)\, s_i^2$$

where $n = \Sigma n_i$ and s_i^2 is the estimated variance of response within the i^{th} treatment. Since, for the binomial samples, we have

$$s_i^2 = P_i(1 - P_i)$$

it follows that

$$MS_w = \frac{1}{n - k} \sum_{i=1}^{k} (n_i - 1)\, P_i\, (1 - P_i)$$

Under Scheffe's method, we may simultaneously test for differences in failure rate among <u>any</u> number of engine family pairs (i, j) at significance level α by checking for whether the inequality

$$|P_i - P_j| > \left[(k - 1)F_{1-\alpha}(k - 1, n - k) \cdot MS_w\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{\frac{1}{2}}$$

is satisfied, where $F_{1-\alpha}(k - 1, n - k)$ is the $100(1 - \alpha)$ percentile of the F-distribution with $k - 1$, $n - k$ degrees of freedom.* Under the multiple-t method, if we preset the total number of comparison tests we wish to make at m, then the above test changes to checking for satisfaction of

$$|P_i - P_j| > t_{1-\alpha/2m}(n - k) \cdot \left[MS_w\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{\frac{1}{2}}$$

where $t_{1-\alpha/2m}(n - k)$ is the $100(1 - \alpha/2m)$ percentile of the t-distribution with $n - k$ degrees of freedom. Recall, in this application, n is the total number of vehicle tests and k is the number of different engine families to which the vehicles belong. α is at the user's discretion, but typical values used are 0.01 and 0.05. Inasmuch as $n - k$ is expected to be quite large,

$$(k - 1)F_{1-\alpha}(k - 1, n - k) \quad \text{and} \quad t_{1-\alpha/2m}(n - k)$$

may be approximated by

$$\chi^2_{1-\alpha}(l - 1) \quad \text{and} \quad z_{1-\alpha/2m},$$

respectively (referring, in turn, to $100(1 - \alpha)$ percentile and $100(1 - \alpha/2m)$ percentile points of the chi-square and standard normal distributions). Finally, consideration of the likely ranges of interest for k and m lead to the conclusion that the multiple-t method will invariably have the greater power for a fixed level of significance α. Thus, we have reduced the multiple comparison tests of interest to that of checking for satisfaction of

$$|P_i - P_j| > z_{1-\alpha/2m} \cdot \left[MS_w\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{\frac{1}{2}}$$

* Statements about the significance level of this and subsequent tests are only approximate because the individual vehicle responses are clearly not normally distributed with homogeneous variance. However, the approximation is expected to be reasonably good because of the generally large sample sizes.

in order to establish that the true failure rates for the engine families in question, p_i and p_J, can be asserted to be different. As a concrete example, suppose we have 50 engine families, each with 1000 vehicle tests (for a total n of 50,000) and that pooled mean square MS_w is 0.02 (corresponding, roughly, to a mean failure rate of 2%). Select level of significance $\alpha = 0.01$, and assume that, at most, 50 comparison tests will be made ($m = 50$). Then we need to find the 99.99 percentile point of the standard normal distribution, which is 3.72, and this sets the value of the right-hand expression in the above inequality to 0.023. This means that any two engine families whose calculated failure rates differ by more then 2.3% may be inferred to have different true failure rates. (At most $m = 50$ such comparison tests are permitted to keep the level of significance at $\alpha = 0.01$.)

Incidentally, if we want to be able to assert that $p_i > p_J$ then replace the above by the corresponding one-sided test, viz., check for satisfaction of

$$P_i - P_J > z_{1-\alpha/m} \cdot \left[MS_w\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{\frac{1}{2}}$$

Remember, for significance level α to be applicable, m must be a preset maximum number of comparisons we are permitted to make.

The above statistical comparison test will ultimately prove to be useful, but it first requires an externally imposed criterion or line of demarcation to define the meaning of "high failure rate." The following procedure is proposed. Rearrange the engine families in the data set in increasing order of P_i. We will have thereby generated a new sequence $(n_1', P_1'), \ldots, (n_k', P_k')$ with $P_i \leq P_{i+1}$. Select a fraction r for partitioning of the engine families into "normal" and "candidate high rate" families. A typically recommended value for r is 0.6. Find the smallest index ℓ such that

$$\sum_{i=1}^{\ell} n_i' \geq r \sum_{i=1}^{k} n_i' = rn$$

The set $(n_1' P_1'), \ldots, (n_\ell', P_\ell')$ constitutes the defined core set of "normal failure rate" engine families. The remaining $m = k - \ell$ engine families are denoted "candidate high failure rate" families.

An alternative way of establishing the core normal failure rate set might be to select a maximum acceptable failure rate p^* and find the largest index ℓ such that $P_\ell' \leq p^*$. For every r criterion, as defined above, there will be an equivalent p^* criterion that results in the same partition. As an example, $p^* = 6\%$ was imposed on the model year 1982 LDV data set for Arizona (for the 220/1.2% cutpoints failure rates). This partitioned the data set of 122,000 vehicle-tests into a "normal failure rate" set of 102,000 vehicle-tests (with mean failure rate of 2.5%) and a "candidate high failure rate" set of 20,000 vehicle-tests (with mean failure rate of 12.1%). The equivalent r criterion would have been $r \simeq 0.84$.

After establishing the core normal failure rate set, coalesce it into a single pooled sample of size

$$n_0 = \sum_{i=1}^{\ell} n_i'$$

and estimated failure rate

$$P_0 = \frac{1}{n_0} \sum_{i=1}^{\ell} n_i' P_i'$$

The multiple-t method with the one-sided test option is now applied. Recall that there are $m = n - \ell$ candidate high failure rate engine families: $(n_{\ell+1}', P_{\ell+1}'), \ldots, (n_k', P_k')$. Compute MS_w, as previously defined, using the pooled normal failure rate set as a single family or "treatment." Select a desired level of significance. A recommended value is $\alpha = 0.05$. Perform the m one-sided comparison tests:

$$P_i' - P_0 > z_{1-\alpha/m} \cdot \left[MS_w \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right]^{\frac{1}{2}} \quad ; \qquad i = \ell + 1, \ldots, k$$

If the inequality is satisfied, engine family (n_i', P_i') is designated a high failure rate family. If not, the engine family is set aside. After all m comparison tests are completed, the set-aside families are absorbed into the core normal failure rate family and the combined collection referred to as normal failure rate families. This collection may thus include some estimated high rate engine families which could not be asserted with confidence to be high rate families. The net result is to define a final collection of high failure rate engine families.

If the above method were applied to the previously illustrated example of 20 engine family samples, a very plausible outcome, depending on reasonable choice of r or p* and α, could have been that families 15 through 20 would have initially been designated candidate high-failure-rate families, but that, after application of the multiple-t method, only families 18, 19, and 20 would retain the high rate designation.

A comment should be made about the possible role of cluster analysis in finding "natural" partitions of engine families into similar groups or clusters. A comprehensive treatment of this methodology is given by Hartigan.[5] Unfortunately, much of the emphasis is on multidimensional deterministic data. The usual approach is to introduce a metric from which a "distance" can be derived for every pair of data points. The aim of clustering is to minimize intra-cluster distances while maximizing inter-cluster distances. A reasonable distance definition for engine families could be the closest separation between the ±1 standard error intervals centered about their estimated failure rates. Applying this definition to the previous example, the distance from family 5 to 16 would be zero, while from 17 to 19 would be about 5 percentage points. A cluster procedure might then establish 19 and 20 as a single cluster and the remaining families into perhaps one, two, or three clusters. How number 18 fares would depend on the particular algorithm and optimization criterion used. The SAS package has a cluster algorithm which unfortunately uses an internally generated Euclidean distance that cannot be accommodated to provide the interval-separation distance function described above. Some problems with the application of cluster analyses to the present problem are that multiple clusters could evolve that have no useful interpretation and that partitioning may expressly not occur in the region of failure rate values where one would like to see the distinction between normal and high failure rates being made.

A note on EPA's use of a "χ^2 test comparing each family's failure rate to the fleet average failure rate." This phrase does not precisely define the procedure in use. We presume it is the following. Define,

$$\left.\begin{aligned} f_i &= n_i P_i \\ s_i &= n_i(1 - P_i) \end{aligned}\right\} \text{Reconstituting the original numerical counts of failures and passes within the ith family}$$

$$\left.\begin{aligned} q_i &= \Sigma f_J - f_i \\ t_i &= \Sigma s_J - s_i \end{aligned}\right\} \text{Counts of failures and passes within all the other families}$$

$$n = \Sigma n_i$$

$$P = \Sigma f_i / n \qquad \text{Fleet average failure rate}$$

$$\left.\begin{aligned} f_i^E &= n_i P \\ s_i^E &= n_i - f_i^E \\ q_i^E &= (n - n_i)P \\ t_i^E &= n - n_i - q_i^E \end{aligned}\right\} \text{Expected counts of failures and passes assuming homogeneity}$$

For each engine family $i = 1, \ldots, k$, form the 2 x 2 table

	FAIL	PASS
i^{th} Family	f_i	s_i
All the Rest	q_i	t_i

and compute the single-degree of freedom chi-square statistic

$$\chi^2 = \frac{(f_i - f_i^E)^2}{f_i^E} + \frac{(s_i - s_i^E)^2}{s_i^E} + \frac{(g_i - g_i^E)^2}{g_i^E} + \frac{(t_i - t_i^E)^2}{t_i^E}$$

to test for homogeneity, i.e., equality of proportions. If $\chi^2 > \chi^2_{1-\alpha}$ (1), then the engine family's i^{th} true failure rate can be said to differ from the true failure rate of all the rest, at significance level α. If, furthermore, $f_i > g_i$, the statement may be amended to state that the i^{th} engine family's true failure rate is greater than the failure rate of all the rest.

Several problems are seen with this presumed procedure. First, it relies entirely on the notion of statistical significance. It is well known that, given sufficiently large sample sizes, just about all compared populations will be significantly different. Second, it is a multiple comparison test and the significance level needs to be appropriately reduced to maintain overall level α. Third, if a particular engine family is determined to be a high failure rate family at some stage of this sequential procedure, it should subsequently be removed from the total class of engine families. Nevertheless a problem would still remain in that the procedure may then be sensitive to the order in which the comparisons are made.

4.4 QUESTION NO. 2: GAUGING AND ADJUSTING FOR TECHNOLOGY GROUP IMPACT

The question literally asked is: should each family be compared to others in the same technology group? We reformulate the question as follows. Can the technology group character of an engine family be used to explain some of the variations in failure rates among families? If so, can the failure rates be adjusted to remove effects due to the use of different technologies so that remaining differences among families due to other causes would be highlighted?

We trust that this reformulation is sufficiently comprehensive to cover the intent of the original question.

The answers are, of course, yes. In fact, a further generalization of the question is suggested -- why not also look to other characteristics, such as manufacturer, model year, and LDV/LDT class as potential contributing explanatory factors for failure rate variations among engine families? In particular, a cursory examination of the data provided suggests marked systematic influences associated with specific manufacturers. The extension to cover LDV's and LDT's as well as multiple model years would help to provide a more unified framework for interpretation of the data.

We propose an additive linear model (with no interactions) as follows.

Let index $i = 1, \ldots, I$ denote manufacturer (M)

$j = 1, \ldots, J$ denote emission control technology group (T)

$k = 1, \ldots, K$ denote LDV, LDT class, respectively (C) $(K = 2)$

$\ell = 1, \ldots, L$ denote model year (Y)

$m = 1, \ldots, M(i, j, k, \ell)$ denote m^{th} engine family within cell (i, j, k, ℓ) (F)

Define $P_{ijk\ell m}$ to be the observed failure rate of the m^{th} engine family within cell $(i, j, k, \ell,)$. The model is expressed as:

$$P_{ijk\ell m} = \mu + \alpha_i + \beta_i + \gamma_k + \delta_\ell + \theta_{ijk\ell m} + \varepsilon_{ijk\ell m}$$

with constraints,

$$\Sigma\alpha_i = \Sigma\beta_j = \Sigma\gamma_k = \Sigma\delta_\ell = 0$$

and

$$\sum_{m=1}^{M(i,j,k,\ell)} \theta_{i,j,k,\ell,m} = 0 \text{ for all } i, j, k, \ell$$

the parameters p, α_i, β_j, γ_k, δ_ℓ, $\theta_{ijk\ell m}$ which represent overall mean, M effects, T effects, C effects, Y effects, and E effects, respectively, are estimated from the observed $\{P_{ijk\ell m}\}$ data. The ε terms represent residual (unexplained) effects.

The above proposed model can be readily implemented on available software packages which provide for linear analysis of categorical responses. In particular the SAS package[3] has the FUNCAT procedure which is sufficiently comprehensive to handle the very unbalanced (wide sample size variations) and sparse (not all (i, j, k, ℓ) cells occupied) type of problems which would be characteristics of the $\{P_{ijk\ell m}\}$ data set. For example, for model year 1982 vehicle-tests in Connecticut, we found that among the 9 x 10 x 2 = 180 possible manufacturer x technology x vehicle class cells only 38 were occupied by at least one engine family.

The output of FUNCAT, in addition to parameter estimates, provides chi-square statistics for testing hypotheses that each of the main effects is significant, and that each of the individual parameter estimates is significantly different from zero. It also computes the chi-square statistic for assessing the level of significance of the residual or unexplained effects. Once statistical significance is established for main effects and individual parameters, the issue of substantive significance can be investigated. For example, if M, T, C, and E effects were all found significant, but not Y effects, and if

$p = 3.5\%$	(significant)
$\alpha_1 = 1.3\%$	(significant)
$\beta_1 = -0.7\%$	(not significant)
$\gamma_1 = -2.0\%$	(significant)
$\theta_{11111} = 1.4\%$	(not significant)
$\theta_{11112} = -1.4\%$	(not significant)

one might draw the following conclusions: for both engine families 1 and 2 within the manufacturer-1, technology-1, vehicle class-1 (LDV), model year-1 cell, a reasonable estimate for failure rate is 3.5 + 1.3 - 2.0 = 2.8%; this number is explained as the sum of 3.5 - 2.0 = 1.5% (mean failure rate for all LDV's) and 1.3% (effect due to manufacturer 1). Note that even though T effects are found overall to be significant, the particular estimate for β_1 (representing the contribution due to technology group 1) is statistically not significantly different from zero. Hence, it is treated as a zero contribution. Other technology groups must have had a significant impact in order for the overall technology effect to be significant, but apparently not group 1. A similar argument leads to the neglect of the ±1.4% estimates for the two engine family contributions. On the other hand suppose that θ_{11111} and θ_{11112} were both statistically significant but evaluated at ±0.2%. This possibility could arise if the two families in question had very large sample sizes. In this instance, one could view the individual engine family effect as <u>substantively</u> insignificant and again choose to ignore it, keeping a common failure rate estimate of 2.8% for both families.

Hopefully, a non-interactive effects model will prove to be adequate, as would be evidenced by a small or insignificant level of residual effects. Such a result would lead to relatively simple and plausible explanations for sources of failure rate variation. If residual effects come out to be significant, one might wish to explore certain interactions, but this extension will be limited by the degrees of freedom available in the sparse experimental design for the problem under consideration.

In summary, application of a categorical response linear model to the vehicle-test data would help to identify the major sources for variation in observed failure rates. It would, in effect, also allow each engine family to be compared to others having common features, like same technology group or same manufacturer.

4.5 QUESTION NO. 3: EFFECT OF REDUCED FAILURE RATES ON METHODS

If failure rates among all engine families follow a generally diminishing trend with successive model years, but the desired level of precision remains invariant, then the situation would actually improve. On the other hand, if the required precision is also reduced in direct proportion to the lowered mean overall failure rates, then the situation would worsen. These conclusions derive from the fundamental properties of binomial distributions. Suppose emission tests were performed on n vehicles belonging to an engine family whose underlying probability of failure is p. Then the resulting number of failures F is a binomial random variable with mean np and standard deviation $\sigma_F = \sqrt{np(1 - p)}$. Consequently, the derived failure proportion or failure rate $P = F/n$ has mean $= p$ and standard deviation $\sigma_p = \sqrt{p(1 - p)/n}$. Our principal interest is in $p \ll 1$ so that $\sigma_p \cong \sqrt{p/n}$. Thus, we observe that, on an absolute scale, σ_p decreases with decreasing p whereas, on a relative scale, $\sigma_p/p \cong 1/\sqrt{pn}$ increases with decreasing p.

To illustrate, if $n = 1000$ and $p = 3\%$, $\sigma_p = 0.54\%$. A reasonable measure of precision might be $2\sigma_p$, which then yields 1.10% on an absolute basis; on a relative basis the precision is 37% of the mean. If p were to reduce to 1%, absolute precision would improve by dropping in value to 0.63%. On the other hand, relative precision would deteriorate to 63% of the mean.

As far as the previously described statistical procedures are concerned, they would continue to be applied in the same way. The methodology itself is not dependent on the actual values of the underlying failure rates. However, the power of the procedures, i.e., their ability to reach significant conclusions, may be affected. If we assume that as individual engine family failure rates follow the downward trend, the separations of p values among families also diminish in proportion, i.e., the phenomenon is likened to a general contraction in scale, then relative precision would be the proper measure to apply. What would then happen in the multiple-t test for high rate families is that a particular family which is truly high rate but does not have a sufficiently high sample size is more likely to be outside the rejection region of the statistical test, i.e., be set aside and not designated a high rate family. Similarly, in application of the categorical response linear model, true effects due to technology, etc. that are of marginal intensity are more likely to be classified as not significant.

It is not necessarily true however that downward trends in failure rates are describable by a general scale contraction. It is entirely possible for <u>some</u> of the failure reductions to follow a simple scale translation rather than a contraction, in which case discriminability should actually increase. However, not all rate reduction can be translations because rates can't reduce below zero.

The net conclusion which we draw is that the statistical procedures, themselves, need not be modified, but that the power of these methods will likely (though not necessarily) be reduced. If power does reduce, a compensatory strategy is to increase n, i.e., accumulate more data. This could be accomplished by using a longer time interval; for example, waiting for six months of data where the previous practice was to commence analysis at three months.

4.6 QUESTION NO. 4: COMBINING DATA ACROSS STATES

State-specific data are easily incorporated into the categorical response linear model described under Question 2 and for which a packaged procedure is readily available within SAS. In detail, we introduce

index n = 1, ..., N denotes state (S)

and revise the model as follows:

$$p_{ijk\ell mn} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + \theta_{ijk\ell m} + \lambda_n + \epsilon_{ijk\ell mn}$$

with the added constraint, $\Sigma\lambda_n = 0$. The parameters λ_n represent the effects of individual states on failure rates. As before, only a main effect is introduced in the anticipation that there are no appreciable interactions with other factors. Note that an "n" index is not added to the engine family parameters $\theta_{ijk\ell m}$ because engine families identify across states.

Assuming that noninteractive state effects will be found adequate, the augmentation of the model should have two significant benefits. First, it would permit one to derive more valid measures of state influences than could be inferred from simple comparison of overall state data set means. The reason, obviously, is that state-to-state differences in the detailed distributions of other effects (technologies, manufacturers, etc.) introduce spurious differences in the raw means. The second benefit is that the additional multistate data should add power to the determination of effects due to the other factors. Of course if strong interactions between state and other factors were to be demonstrated by a large increase in the residual variation, then these benefits may diminish or be vitiated. As noted before, limited exploration of interactions can be conducted, if necessary.

4.7 QUESTION NO. 5: HOW TO ANALYZE RESULTS FROM ALTERNATIVE SETS OF CUTPOINTS

Since EPA is interested in alternative sets of cutpoints, the recommended statistical tests should be done for each set. High failure rate engine families can be identified in each set, but not necessarily using the same external standard or line of demarcation between "normal" and "high" rates. In fact, cursory examination of some of the data suggests that the 100/0.5% criterion results, on the average, in roughly three times the failure rate as that produced by the 220/1.2% criterion. If the r fraction method were used for set partitioning, it would tend to naturally establish a higher equivalent failure rate criterion for the 100/0.5% cutpoints. This may be an argument for using the r fraction method rather than the direct p^* criterion since the latter requires separate designation of p^* for the two sets of cutpoints. The results may not then be comparable in severity of the pruning achieved.

Because of the substantially higher failure rates associated with the more stringent cutpoints (which can be viewed as a scale expansion effect), we should expect increased relative precision and therefore sharper ability (via the multiple-t tests) to determine that a candidate high failure rate engine family is a high rate family when it truly is in that category.

The FUNCAT procedure in the SAS package readily accepts categorical responses of any multiplicity (and even dimensionality). Thus, for each vehicle-test sample one would read in, in addition to the design effects categories (i, j, k, ℓ, m, n), the response data as n, P_1 and P_2 rather than just n, P. There is provision within FUNCAT to define a scalar response function of the input probabilities. One could first select P_1, run the model, then select P_2 and rerun the model to get an analysis of variance and significant effects estimation with reference to each set of cutpoints. Comparison of the two results may shed light on the sensitivity of various effects to cutpoint criteria. These have to do with the detailed distributions of measured HC and CO concentrations within individual engine families. If these distributions are fairly smooth and similar in shape (in the vicinity of the cutpoints) over most engine families, then one should not expect much difference in effects evaluation for the two cutpoint sets. Suppose the distribution saturates between HC cutpoints for some families but not for others, depending, say, on technology type, then profound differences in significant effects may be found in the two analyses.

The flexibility of FUNCAT with respect to response function also permits running the model for such response combinations as $P_1 - P_2$ or P_1/P_2. These results would help focus on effects which contribute to translational or scaling dissimilarities over the set of engine families.

REFERENCES FOR SECTION 4

1. Analysis of Emission Test Failure Rates in Centralized I/M Programs, Report to the EPA, EEA, September 1986.

2. Kramer, C. Y., Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications, Biometrics, 12, 307-310, 1956.

3. SAS User's Guide, 1979 Edition, SAS Institute Inc., Cary, NC, 1979.

4. Afifi, A. A. and Azen, S. P., Statistical Analysis - A Computer Oriented Approach, Academic Press, New York, 1972, pp 74-75.

5. Hartigan, J. A., Clustering Algorithms, John Wiley & Sons, New York, 1975.

5. SUMMARY AND CONCLUSIONS

The identification of "pattern case" failures using emission data from centralized I/M programs for 1981 and later model year light-duty vehicles and light-duty trucks has been investigated in this work assignment. The objectives are to define a system that will lend itself to rapid, periodic analysis of data. Our investigation identified issues in three areas:

- Selection of I/M programs from which one can obtain relatively clean data for the analysis with little time lag.
- A simplified processing scheme to minimize costs and turnaround time.
- Improved statistical methods to better define pattern failures and possibly identify effects of test procedures and/or ambient variables.

The selection of I/M programs for data analysis depends, to some degree, on the questions being investigated. Our analysis shows that highly automated programs such as those in Illinois, Wisconsin and Kentucky are best suited in terms of data cleanliness and rapid "turnaround" of test data. It will be possible to perform analysis on a quarterly basis if data form these states are used. Unfortunately, these states also utilize different test procedures, preventing easy data comparison across states. Earlier analysis of data suggests that EPA should examine data from states using identical test procedures and across states using different test procedures, so that false failures related to the test procedures can be identified. Moreover, all centralized programs are improving their data acquisition methods. In a few years, data from all states may be relatively similar as far as cleanliness and turnaround time. At the current time, we would recommend the following

- Investigate the three procedures currently used - idle with no preconditioning, idle with 2500 rpm preconditioning, and idle with loaded-mode preconditioning.

- Investigate I/M programs such that each test procedure type is utilized in at least two geographically distinct programs. This will require investigation of six I/M programs, at least.
- Select the six I/M programs from the universe of I/M programs based on a combination of sample size (at least a total sample of 50,000 per month), automated data acquisition and rapid turnaround. We believe that realistically, analysis of data bi-annually will be possible from six programs.
- California engine families require California data. Although it has several drawbacks, the California data also suggests some interesting possibilities that may make an engine family specific analysis feasible.

Data processing steps required after acquiring the data from the states include cleaning, sorting, VIN decoding and failure rate calculations. The cleaning step is a general "front-end" step that will require highly variable efforts, depending on the relative cleanliness of the input data. However, even the best data sources require some cleaning, if only to eliminate calibration, aborted tests, heavy-duty vehicles, etc. One of the problems is that each state's program is constantly being changed and the cleaning steps will have to reflect these changes This step, therefore, requires programmer intervention and can be tedious

Sorting and sequencing of data is required to recognize initial tests and retests for the same vehicle Although other elaborate schemes have been considered, a VIN based sort may be adequate for this analysis Sequencing is not an issue if pattern cases are to be recognized based only on a first test failure; other issues, such as vehicle repairability and waiver rates may be interesting to EPA but will require additional sequencing steps.

Other steps including VIN decoding and calculation of failure rates are straightforward. Identification of transmission type is generally not possible. In addition, there is no easy resolution to the "running change in certification" problems. Problems with end of model year (MYR) vehicles and certification family MYR versus vehicle MYR appear to be restricted to very few cars. EPA certification staff have claimed that proper specification of carryover engine families should be no problem if the final version of the certification tape is used; there may be some residual problems unknown to EPA.

Once failure rates by engine families have been calculated, a number of statistical tools can be employed to identify pattern cases and address several related issues. We recommend a test called Scheffe's multiple-t test to both define and identify pattern failures, and this test is an improvement over EPA's current "X^2" test. We have proposed statistical linear models that can evaluate technology specific, manufacturer specific, and testing procedure specific influences. In addition, we have suggested methods to use data from two sets of cutpoints, and methods to combine data from several states. All of the proposed methods fit well into the processing framework, in that they are available on SAS (Statistical Analysis System).

EPA has inquired about sample size requirements for the analysis, and this <u>cannot</u> be answered in the absolute sense The sample size required to identify any particular engine family depends on:

- The sales of that engine family
- The failure rate of that family in comparison to the fleet average failure rate
- The statistical significance with which EPA can claim the family is a "pattern case"
- Convoluting factors such as technology specific rates and response to ambient conditions.

At this point, it does not serve any purpose to fix a sample size; rather, as sample sizes are increased, one can expect pattern failures to be recognized for low sales families with greater precision.

Finally, EPA has requested some specific estimates on cost. We have attempted to estimate a cost for analysis of one years' worth of data form a program which tests 100,000 vehicles per month, and one-third of the vehicles are from the five newest model years. Thus a total of 1.2 million vehicle inspections (up to 1.6 ~ 1.8 million records) are obtained, and 400,000 vehicles' data are separated, cleaned, sorted and VIN decoded. The resulting output of failure rates by engine family are then statistically tested for pattern cases, no other statistical analysis is performed. Costs are summarized in Table 5-1 for *one* such data source. As can be seen, computer costs, if the analysis is done on a private time-sharing mainframe, are very high. On the other hand, access to government computers can reduce computer costs by a factors of 3. Costs for analysis of six programs will be six times the estimate; however, the estimate does *not* scale linearly with sample size. Halving the sample size will reduce costs only by about 25 percent.

TABLE 5-1

COST OF DATA ANALYSIS

Assumptions - Initial tape has 1 2 million vehicles of which 400,000 are 1981+ light-duty. Processing on mainframe - IBM 3033 or equivalent.

CPU time (initial cleanup)[a]	60 minutes
CPU time (all other processing)	120 minutes

Computer Costs (government system)

CPU time @ $1/sec	$10,800
I/O costs	$1,000
Tape storage	$250
Disk storage	$250
Connect time	$700
Total	$13,000

Computer costs; Private systems	~$40,000

Labor costs

Programmer	120 hours
Manager	40 hours
Analyst	40 hours
Total cost @ $45.00/hour	$9,000

[a] May be higher or lower depending on data source