The papers that follow sketch the content of 60 publications, of which 7 are books. It also covers 8 technical reports and 13 internal working papers, as well as considerable amounts of unreported and unpublished material.

Some of the more innovative contributions brought to full exposition include:

1) A volume on Configural Polysampling, so far the only direct approach to robustness in finite-sized samples.
2) A volume on Exploratory Analysis of Variance, taking an approach which is both novel and effective.

Other important innovations include:

3) New light on multiple-comparison problems, especially as they arise in the analysis of variance.
4) Use of simple regression (on an orthogonal space) as a means of composite building.
5) Development of general techniques of shape comparison
6) Progress on the "separations problem".
7) Discussion of how resampling methods (jackknife, bootstrap) should be applied to problems where blocking (in the design of experiments sense) is essential.
8) Use of limited lateral randomization in visualizing distributions.
9) Introduction of novel, more effective measures of urbanization.
10) Development of new, apparently promising approaches to clustering.

4. SUBJECT TERMS
Data analysis, Statistics, Robustness, Analysis of variance, Multiple comparisons, Shape comparison, Resampling with blocking, Limited lateral randomization

15. NUMBER OF PAGES
17
16. PRICE CODE
UL
Data Analysis

by

John W. Tukey

FINAL REPORT
May 1, 1986 — December 31, 1990
DAAL03-86-K-0073

supported by the

U. S. Army Research Office (Durham)
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>1</td>
</tr>
<tr>
<td>1. Personnel 1986 — 1990</td>
<td>2</td>
</tr>
<tr>
<td>Books, chapters in books</td>
<td>2</td>
</tr>
<tr>
<td>Papers</td>
<td>6</td>
</tr>
<tr>
<td>Submitted for publication:</td>
<td>7</td>
</tr>
<tr>
<td>Papers delivered and in preparation:</td>
<td>7</td>
</tr>
<tr>
<td>3. Theses</td>
<td>8</td>
</tr>
<tr>
<td>Technical Reports: Department of Statistics</td>
<td>8</td>
</tr>
<tr>
<td>Princeton University</td>
<td>8</td>
</tr>
<tr>
<td>Technical Reports: Temporary series -- Department Statistics</td>
<td>9</td>
</tr>
<tr>
<td>Princeton University</td>
<td>9</td>
</tr>
<tr>
<td>Technical Reports: Department of Civil Engineering</td>
<td>9</td>
</tr>
<tr>
<td>Princeton University</td>
<td>9</td>
</tr>
<tr>
<td>Internal Working Papers (IWP)</td>
<td>10</td>
</tr>
<tr>
<td>Statistical software</td>
<td>10</td>
</tr>
<tr>
<td>5. Sketch of work May 1986 — 1990</td>
<td>11</td>
</tr>
<tr>
<td>6. Access</td>
<td>11</td>
</tr>
<tr>
<td>7. Analysis of Variance (ANOVA)</td>
<td>12</td>
</tr>
<tr>
<td>8. Clustering</td>
<td>13</td>
</tr>
<tr>
<td>9. Graphical Techniques</td>
<td>13</td>
</tr>
<tr>
<td>10. Hints</td>
<td>13</td>
</tr>
<tr>
<td>11. Multiple comparisons</td>
<td>14</td>
</tr>
<tr>
<td>12. Randomization</td>
<td>14</td>
</tr>
<tr>
<td>13. Regression (and related matters)</td>
<td>15</td>
</tr>
<tr>
<td>14. Robustness</td>
<td>15</td>
</tr>
<tr>
<td>15. Shape</td>
<td>15</td>
</tr>
<tr>
<td>16. Smoothing</td>
<td>16</td>
</tr>
<tr>
<td>17. Stability of results</td>
<td>17</td>
</tr>
<tr>
<td>18. Techniques, computational</td>
<td>17</td>
</tr>
<tr>
<td>19. Techniques, statistical</td>
<td>18</td>
</tr>
<tr>
<td>20. Urbanization measures</td>
<td>18</td>
</tr>
</tbody>
</table>
Executive Summary

The papers that follow sketch the content of 62 publications, of which 7 are books. It also covers 12 technical reports and 13 internal working papers, as well as considerable amounts of unreported and unpublished material.

Some of the more innovative contributions brought to full exposition include:

1) A volume on Configural Polysampling, so far the only direct approach to robustness in finite-sized samples (Tukey 1991c).

2) A volume on Exploratory Analysis of Variance, taking an approach which is both novel and effective (Tukey 1991h).

Other important innovations include:

3) New light on multiple-comparison problems, especially as they arise in the analysis of variance (Tukey 1991n).

4) Use of simple regressions on orthogonal space as a means of composite building where the data is not strong enough for conventional multivariate techniques (Tukey 1991o).

5) Development of general techniques of shape comparison that take advantage of such available machinery as weighted least squares and conventional multivariate analysis (Goodall, see Section 15, below).

6) Progress on the "separations problem" where we ask if a batch of numbers is better thought of as two (or more) subbatches (Technical Reports 293, 298).

7) Discussion of how resampling methods (jackknife, bootstrap) should be applied to problems where blocking (in the design of experiment sense) is essential (Technical Report 292).

8) Use of limited lateral randomization in visualizing distributions involving many (100 to 10,000) points (Tukey and Tukey 1991Se).

9) Introduction of novel, more effective measures of urbanization (Kafadar and Tukey 1990Sa).

10) Development of new, apparently promising approaches to clustering (Hansen & Tukey 1990Sb).

John W. Tukey
Princeton, 2 May 1991
1. Personnel 1986 — 1990

Faculty

Colin Goodall 1986-87
John W. Tukey 1986-90

Visiting Faculty (short term)

Thu Hoang April-June 1987, February 1988, March 1989
Catherine Marsh September 1988
Karen Kafadar December 1989
Kaye Basford January-February 1990

Graduate Students

Ha Nguyen 1986 (Ph D. 1986)
Katherine M. Hansen (Ph.D. 1989)

Undergraduate Student

Bill Frack August-September 1987

Research Assistant

E. Olszewski 1986-90


Books, chapters in books:


NOTE: Letters used with years on John Tukey's papers correspond to bibliographies in all volumes of his collected papers.


Papers:


**Submitted for publication:**


**Papers delivered and in preparation:**


Tukey, John W. (1990Ug). "A suggested, more unified approach to multiplicity".


3. Theses

Ph.D. Thesis

1986—


1989—

Hansen, Katherine M., "Some statistical problems in geophysics and structural geology," June.


Technical Reports: Department of Statistics, Princeton University

<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Author and date</th>
</tr>
</thead>
<tbody>
<tr>
<td>291</td>
<td>Thinking about non-linear smoothers</td>
<td>John W. Tukey</td>
</tr>
<tr>
<td></td>
<td></td>
<td>May 1986</td>
</tr>
<tr>
<td>292</td>
<td>Kinds of bootstraps and kinds of jackknives,</td>
<td>John W. Tukey</td>
</tr>
<tr>
<td></td>
<td>discussed in terms of a year of</td>
<td>April 1987</td>
</tr>
<tr>
<td></td>
<td>weather-related data</td>
<td></td>
</tr>
<tr>
<td>293</td>
<td>Procedures for separations within batches of</td>
<td>Thu Hoang</td>
</tr>
<tr>
<td></td>
<td>values, I. The orderly tool kit and some</td>
<td>John W. Tukey</td>
</tr>
<tr>
<td></td>
<td>heuristics</td>
<td>March 1989</td>
</tr>
<tr>
<td>294</td>
<td>Tuning a major part of a clustering algorithm</td>
<td>Katherine M. Hansen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>John W. Tukey</td>
</tr>
<tr>
<td></td>
<td></td>
<td>February 1988</td>
</tr>
<tr>
<td>Number</td>
<td>Title</td>
<td>Author and date</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>57</td>
<td>Principal components analysis of neural and facial skull configurations as a measurement of orthoccephalization in ralte</td>
<td>C. R. Goodall, A. Bose, G. Das Kapta, February 1986</td>
</tr>
<tr>
<td>58</td>
<td>Change-of-shape: a production system of $S$ macros for growth analysis</td>
<td>C. R. Goodall, March 1986</td>
</tr>
<tr>
<td>60</td>
<td>Characterization of skew-co-ordinate duality</td>
<td>C. R. Goodall, May 1986</td>
</tr>
</tbody>
</table>

Technical Reports: Department of Civil Engineering, Princeton University

<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Author and date</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOR 87-9</td>
<td>Interpolation of multivariate data</td>
<td>Colin R. Goodall, M. Thoma, November 1987</td>
</tr>
<tr>
<td>SOR 87-11</td>
<td>The use of robust methods for shape comparisons</td>
<td>Colin R. Goodall</td>
</tr>
<tr>
<td>SOR 88-7</td>
<td>The analysis of averages and the analysis of variance</td>
<td>Colin R. Goodall, April 1988</td>
</tr>
</tbody>
</table>
### Internal Working Papers (IWP)

<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Author and date</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWP-71</td>
<td>Resistant fitting of quadratics to seven equally spaced points of which some may be missing</td>
<td>John W. Tukey 1987</td>
</tr>
<tr>
<td>IWP-72</td>
<td>Diagnostic tools for character clouds</td>
<td>John W. Tukey 1987</td>
</tr>
<tr>
<td>IWP-73</td>
<td>Comparing empirical distributions over time</td>
<td>John W. Tukey 1987</td>
</tr>
<tr>
<td>IWP-74</td>
<td>Some questions about categorical regression</td>
<td>John W. Tukey 1987</td>
</tr>
<tr>
<td>IWP-75</td>
<td>Three-word access with branching lengths</td>
<td>John W. Tukey 1987</td>
</tr>
<tr>
<td>IWP-81</td>
<td>Validating detectable differences</td>
<td>John W. Tukey 1988</td>
</tr>
<tr>
<td>IWP89-1</td>
<td>Percentage points of the range</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP89-2</td>
<td>Empirical improvement of HMM word-classing schemes</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP89-7</td>
<td>Introduction to paragrammar</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP89-8</td>
<td>Borrowing strength applied to 2-way tables of jackknifed variances</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP89-9</td>
<td>Higher criticism for individual significances in several tables or parts of tables</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP89-10</td>
<td>Class of accumulation patterns useful in support of certain agglomerative clustering algorithms</td>
<td>John W. Tukey 1989</td>
</tr>
<tr>
<td>IWP90-1</td>
<td>A collection of points relevant to making regression incisive</td>
<td>John W. Tukey 1990</td>
</tr>
</tbody>
</table>

### Statistical software

**Goodall, C. R. (1989)**

*ANOVA:* S functions for classical and resistant analysis of averages and the analysis of variance using the sweep operator. S library archive, statlib@temper.stat.smu.edu
5. Sketch of work May 1986 — 1990

Over this period work was carried out on a considerable variety of topics, most directed toward improving the analysis of data. In the next 15 sections, this work is summarized and related to papers and reports under 15 headings: access, anova, clustering, graphical techniques, hints, MC, randomization, regression, robustness, shape, smoothing, stability, techniques (computational and statistical), and urbanization.

Overviews and collected volumes are now discussed first.

The interface between data analysis and computing was reviewed briefly (Tukey 1986k). The impact of the geophysical sciences on statistics and data analysis has been considered (Tukey 1989Ue).

The likely evolution of the data analytic, and statistical, techniques likely to be used by the members of the American Statistical Association's Section on Physical and Engineering Sciences -- which will reflect quite well the techniques used across much broader areas of application -- was forecast and the relevance of describing the evolution of each of many data analytic techniques in terms of three consecutive 30-year periods was pointed out. (Tukey 1989w)

An update of "Future of Data Analysis (originally published by Tukey in (1962) was requested for an international meeting in France. A partial update, by Morgenthaler and Tukey (1990r), was prepared, presented, published. A fuller version is to be prepared, and is likely to be published in book form.

Volumes III to VI of Tukey's Collected Papers were issued, including 24 previously unpublished papers (see Publications, above).

6. Access

A variety of special topics related to access of full-text documents by searching full text have been explored (Internal Working Papers 75, 89-2, 89-7, 89-10).
7. Analysis of Variance (ANOVA)

Work on co-editing, and writing several chapters for, a book to be called

*Fundamentals of Exploratory Analysis of Variance* continued during the large part of the period being reviewed. Appearance of the book is planned for Fall 1991 (co-editors: David C. Hoaglin, Frederick Mosteller, John W. Tukey, 1991h). This book takes a much more modern - and much more realistic - view of the analysis of variance than anything in print. At least one succeeding volume is in preparation.

Some significant innovations include:

- a more general - more widely applicable - basis for the "Rule of 2" in downsweeping combining some packets (lines) in an initial analysis with each other,

- serious thought about which comparisons in a 2-way table deserve special attention - mainly "submaineffects" and "double differences" (2-way differences exhibiting interaction or its absence),

- use of biranges (= maximum size of double differences in a 2-way table) by analogy with ranges (= maximum size of differences in a 1-way table),

- simple approximations to birange % points,

- extensions to 3-way tables (only discussed lightly).

(More information about authored or coauthored chapters can be found under Publications, above). Some aspects of this work were summarized in (Tukey 1989Ud), others are alluded to in (Tukey 1991Uh).

The application of borrowing strength by median polish in the special case where the tabulated values are jackknived estimates of variance has been considered (TWP89-8, Tukey).

Other work will contribute to the second volume of this series including (Goodall and De Veau 1991U).
8. Clustering

The stage-by-stage development of a major portion of a clustering algorithm has been documented, submitted for publication, and revised (Hansen and Tukey 1990Sb, based on Technical Report 294). The resulting algorithm is quite novel combining a variety of quite distinct subalgorithms and, while it makes no explicit use of a Gaussian distributional assumption, it shows performance against a Gaussianity-distributed test bed that is almost as good as that provided by a Gaussian-likelihood-based algorithm. Thus its performance on real-world not-exactly-Gaussian data may well be better than any of the many algorithms presently available. It is notable that its development and evolution involve clusters that overlap one another seriously.

9. Graphical Techniques

Techniques for displaying distributions, mainly in terms of individual points of a sample have been explored, and innovative possibilities expounded (Tukey and Tukey 1990Se).

Diagnostic tools for examining character clouds have been discussed (IWP-72, Tukey).

Displaying linked aspects of data points has been reviewed and discussed (a technical report begun here will be reported under DAAL03-88-K-0045).

Delineation plots for bivariate data have been developed and discussed (Goodall, Stoughton and Easton, 1986).

(Work after May 1, 1988 in this area is reported under DAAL03-88-K-0045.)

10. Hints

Exploratory data analysis can only serve its functions by detecting and mentioning phenomena, not all of which meet the usual standards (significance, confidence) of
confirmatory data analysis. But mentioning anything and everything dredged up in an extensive and deep exploration is equally unlikely to be helpful. Some guidance for the choice of what is to be mentioned is probably essential.

Catherine Marsi and John Tukey have been considering this problem for at least three years (since 1988) and draft discussions of what to use and how to use it are approaching readiness for publication. (An earlier version is Tukey 1990Ug).

11. Multiple Comparisons

A review paper on the Philosophy of Multiple Comparisons, originally a Miller lecture at Stanford (Tukey 1991n). This paper introduces - - and discusses - - a variety of issues of importance for multiple comparisons. It interrelates substantially with the work on analysis of variance (see Section 7, above) and the work on hints (see Section 10).

An application of the "higher criticism" to the question - - what fractions of individually significant results, when all candidates are divided into bundles (perhaps one bundle for each of several tables), are likely to be real - - has been prepared (IWP89-9, Tukey).

The problem of approximating the distribution of the studentized birange (see Section 7, above) by a well-chosen studentized range distribution has been studied and discussed (Tukey 1989Uc).

12. Randomization

The state of the art of rerandomization as offering an almost completely trustworthy analysis of randomized experiments or data collections, as well as a comparatively highly trustworthy analysis of other data sets has been reviewed and extended (Tukey 1989Ub).
13. Regression (and related matters)

The problem of simple robust regression, in the face of both smoothly-varying variability and exotic values requiring robust estimation has been studied, and a substantial paper will appear (Cohen, Dalal, and Tukey 1991Sa).

The use of many covariates in analyzing timing-of-events experiments has been re-examined and new, effective techniques proposed (Tukey 1991o). Similar approaches should be effective in a wide variety of regression or regression-related circumstances.

Some questions about categorical regression have been considered (IWP74, Tukey).

14. Robustness

Earlier and continuing work on configural polysampling -- a realistic approach to optimum robustness -- has culminated in the appearance (Spring 1991) of a small book (Morgenthaler and Tukey 1991c).

Since a diverse set of techniques for robustly smoothing numerical sequences are now available, it is important to learn how to think about robust smoothers, and particularly about how to select a robust smoother for a particular purpose. These questions have been examined in some depth (Technical Report 291, Tukey). Some ways to make regression more incisive have been discussed (IWP90-1, Tukey).

The resistant fitting of straight lines to 9 or fewer points has been considered (IWP-71, Tukey)

15. Shape

Major emphases in this area include:

- the extension of Procrustes techniques, both least-square and robust, to the comparison of more than two (geometrical) forms (generalized Procrustes techniques)
16. Smoothing

The robust smoothing of sequences, see Section 13 above (Technical Report 291).

Robust smoothing in the plane has received continuing attention. Important ideas include:

- an inverse convex-hull procedure for computing a polygon (or set of nested polygons) surrounding each data point,
- a convenient data structure for computing a median,
- adaptation of the end-value rule for boundary data.

Work on this topic continues.
17. Stability of results

The stability of adjusted (specifically age-adjusted) rates has been discussed (Tukey 1986u).

The use of resampling techniques - jackknife or bootstrap - to assess stability of results of data analysis in those situations where blocking is essential have been examined, and reasonable techniques for doing this have been discussed. (Technical Report 292, Tukey).

Techniques for deciding when it is desirable to discuss a batch of numerical values as two or more subbatches -- solely on the basis of the numerical values themselves -- have been examined. The first results are available as Technical Reports (293 and 298, Hoang and Tukey).

The g-and-h distributions form a useful 2-parameter family, accommodating skewness and elongation. Because they can be fitted in terms of quantiles (order statistics) they may prove considerably easier to estimate than families for which moment estimation seems natural (which turn out to decimal very large sample sizes). Empirical bounds for quantile-based estimates of g have been studied (Hoaglin and Tukey 1989Sc).

The degree to which the adequacy of an attempt to design an experiment of prescribed power can be assessed after the data has been collected has been discussed (IWP-81, Tukey).

The comparison of an ordered set of parallel distributions has been considered (IWP73, Tukey).

18. Techniques, computational

Interpolation in the plane, and in higher dimensions, has been studied, providing a common framework for data interpolation (related to key-frame interpolation) and view interpolation (related to kinematic displays of high dimensional data) (Goodall and Thoma

Functions useful in the analysis of averages have been coded and reported (Goodall, Technical Report SOR 88-7).

Software for the superimposition of forms has been prepeared.

Empirical formulas for unusual % points of the range have been prepared (IWP89-1, Tukey).

19. Techniques, statistical

A convenient, quite detailed table of the distribution of Student's $t$ has been prepared and published (Kafadar and Tukey 1988h).

Techniques for re-expressing exponentially distributed quantities, using simple hybrid re-expressions have been studied and will appear shortly (Tukey 1991p).

20. Urbanization measures

Novel but simple measures of urbanization for geographical units like counties ranging from:

- the logarithm of the size of the largest place


- the logarithm of the square root of the sum of the squares of the sizes of all places

have been tried out in such contexts as cancer rates, (age specific) birth rates, and median family incomes (Kafadar and Tukey 1990Sd, Goodall, Kafadar and Tukey 1990Ua).