The Development of Statistical Design and Analysis Concepts at Rothamsted

Roger Payne,
VSN International, Hemel Hempstead, Herts, UK
(and Rothamsted Research, Harpenden, Herts UK).

email: Roger.Payne@vsni.co.uk
100 years ago

• significance test (Arbuthnot 1710)
• Bayes theorem (Bayes 1763)
• least squares (Gauss 1809, Legendre 1805)
• central limit theorem (Laplace 1812)
  • distributions of large-samples tend to become Normal
• "Biometric School" Karl Pearson, University College
  • correlation, chi-square, method of moments
• t-test (Gosset 1908)
• Fisher in *Stats Methods for Research Workers*
  • "..traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow!"
Rothamsted

- Broadbalk – set up by Sir John Lawes in 1843 to study the effects of inorganic fertilisers on crop yields
- had some traces of factorial structure, but no replication, no randomization and no blocking
- and not analysed statistically until 1919..
Broadbalk

- treatments on the strips
  01 (Fym) N4  11 N4 P Mg
  21 Fym N3  12 N1+3+1 (P) K2 Mg2
  22 Fym  13 N4 P K
  03 Nil  14 N4 P K* (Mg*)
  05 (P) K Mg  15 N5 (P) K Mg
  06 N1 (P) K Mg  16 N6 (P) K Mg
  07 N2 (P) K Mg  17 N1+4+1 P K Mg
  08 N3 (P) K Mg  18 N1+2+1 P K Mg
  09 N4 (P) K Mg  19 N1+1+1 K Mg
  10 N4  20 N4 K Mg

- hints of factorial structure (e.g. plots 03, 05, 10, 09)
- but no replication (nor randomization)

..
Ronald Aylmer Fisher

• in 1919 Rothamsted Director John Russell was faced with "great files of records"
  • looked for a "young mathematician ... prepared to examine our data and elicit further information"
• obtained a reference from Fisher's tutor at Caius College Cambridge
  • The answer was that he could have been a first class mathematician had he "stuck to the ropes" but he would not.
• conclusion
  • That looked like the type of man we wanted, so I invited him to join us. I had only £200, and suggested he stay as long as he thought that should suffice. ... It took me a very short time to realise that he was more than a man of great ability, he was in fact a genius who must be retained. So I set about obtaining the necessary grant.
• Fisher was at Rothamsted from 1919-1935

• reference: Joan Fisher Box (1978) R.A. Fisher The Life of a Scientist
William Sealy Gosset

- worked at Guinness in Dublin
  - published under the pseudonym "Student"
  - devised the t-test (Biometrika 1908)
- approved of Fisher's appointment at Rothamsted
  - "there should be lots of interesting work to be done there and they might easily have got someone there who would have been worse than useless"
- had a long collaboration and correspondence with Fisher
  - .. but when I come to "Evidently" I know that means at least two hours hard work before I can see why (see J.F. Box, p.115)
  ..

Photo from the MacTutor History of Mathematics archive (John J O'Connor and Edmund F Robertson) http://www-groups.dcs.st-and.ac.uk/~history/
t-test

- originally derived empirically by Gosset
  - test for small samples that sample mean $m = \mu$
  - if we use sample estimate of s.e. instead of $\sigma$
    cannot assume $m - \mu$ will still have a Normal distribution
- caused Fisher to recognise concept of degrees of freedom
  - estimate s.e. by $\sqrt{(x - m)^2 / (n - 1)}$ with divisor $n - 1$ not $n$
  gave full mathematical justification & generalized to
  - differences between means
  - coefficients in regressions and multiple regressions
  - or any statistic of form Normal/$\sqrt{\text{Chisquare}}$
  - and provided an expansion of the cumulative distribution formula of the distribution in inverse powers of $n$, together with some tables
ANOVA


<table>
<thead>
<tr>
<th>Variation due to</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuring</td>
<td>5</td>
<td>6,158</td>
<td>1231.6</td>
<td>35.09</td>
</tr>
<tr>
<td>Variety</td>
<td>11</td>
<td>2,843</td>
<td>258.5</td>
<td>16.07</td>
</tr>
<tr>
<td>Deviations from summation formula</td>
<td>55</td>
<td>981</td>
<td>17.84</td>
<td>4.22</td>
</tr>
<tr>
<td>Variation between parallel plots</td>
<td>141</td>
<td>1,758</td>
<td>12.47</td>
<td>3.53</td>
</tr>
<tr>
<td>Total</td>
<td>212</td>
<td>11,740</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

In comparing the standard deviations in the last column we may use the fact that 3.63, for example, has the same accuracy as if it had been determined from a sample of 142; the variance of its natural logarithm is therefore \( \frac{1}{2 \times 141} \). Thus, to test if the deviations from the summation formula are significantly greater than would occur by chance, we compare the difference of the logarithms with its standard error, namely \( \sqrt{\frac{1}{42} + \frac{1}{141}} \):

- 12 varieties \( \times 2 \) dung (+,−) \( \times 3 \) fertilizers (basal, sulphate, chloride)
- ignore block structure (half-field / (plot * row))
- fits main effects of variety and manures, and their interaction tested by using approximate Normality of log(variance)
- then fits a multiplicative model (by eigenvalue decompositions)
Analysis of variance

  • "The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic."

• Yates (1937) *Design and Analysis of Factorial Experiments*
  • including Yates' algorithm for analysis of variance – orthogonal designs, one error term

  • non-orthogonal (balanced) designs, several error terms


• Payne (1998). Detection of partial aliasing and partial confounding in generally balanced designs. *Comp. Statistics* ...
Another tea test

- could Dr B. Muriel Bristol tell whether milk had been poured first?
- Chapter II of *The Design of Experiments* (Fisher 1935)
  - used to illustrate concepts of randomization, significance, exact tests
Significance tests

• Arbuthnot (1710) "Divine Providence had intervened in favour of the male sex"
  • probability $1/2^{17}$ of more male than female births in London in 17 consecutive years

• Fisher (1935) *The Design of Experiments* Ch. II
  • Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis

• Yates "those damned stars"

• Payne quoted in Perry (1986) Multiple-comparison procedures a dissenting view. *J. Econ Entomol*
  • .. likened a significance test to the safety net of a tightrope walker: it helps give confidence {that unjustified conclusions are not being drawn from random errors in the data} but should not be part of the act
<table>
<thead>
<tr>
<th>Centre</th>
<th>Mean Yield</th>
<th>Mean response to</th>
<th>Standard Error</th>
<th>Interactions</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>P</td>
<td>K</td>
<td>N x P</td>
</tr>
<tr>
<td>2. Ipswich</td>
<td>5.17</td>
<td>+0.02</td>
<td>-0.17</td>
<td>-0.15</td>
<td>+0.17</td>
</tr>
<tr>
<td>3. Colwick</td>
<td>7.17</td>
<td>+1.53**</td>
<td>+0.36</td>
<td>+0.21</td>
<td>+0.46</td>
</tr>
<tr>
<td>4. Newark</td>
<td>8.29</td>
<td>+0.10</td>
<td>+0.39</td>
<td>+0.20</td>
<td>+0.36</td>
</tr>
<tr>
<td>5. Felstead</td>
<td>9.09</td>
<td>+0.17</td>
<td>+0.26</td>
<td>+0.56</td>
<td>+0.29</td>
</tr>
<tr>
<td>6. Brigg</td>
<td>10.74</td>
<td>+1.76**</td>
<td>-0.74</td>
<td>+0.10</td>
<td>+0.38</td>
</tr>
<tr>
<td>7. Poppleton</td>
<td>11.71</td>
<td>+0.02</td>
<td>-0.02</td>
<td>+0.04</td>
<td>+0.38</td>
</tr>
<tr>
<td>8. Bardney</td>
<td>12.32</td>
<td>+1.26</td>
<td>+0.01</td>
<td>+0.30</td>
<td>+0.71</td>
</tr>
<tr>
<td>9. Allscott</td>
<td>12.38</td>
<td>+0.38</td>
<td>+0.16</td>
<td>+1.16**</td>
<td>+0.28</td>
</tr>
<tr>
<td>10. Wissington</td>
<td>13.82</td>
<td>+0.20</td>
<td>+0.08</td>
<td>-0.39</td>
<td>+0.28</td>
</tr>
<tr>
<td>11. Peterborough</td>
<td>14.06</td>
<td>-0.29</td>
<td>-0.03</td>
<td>-0.03</td>
<td>+0.26</td>
</tr>
<tr>
<td>12. King’s Lynn</td>
<td>14.36</td>
<td>-0.39</td>
<td>-0.33</td>
<td>-0.65</td>
<td>+0.33</td>
</tr>
<tr>
<td>13. Ely</td>
<td>14.74</td>
<td>+0.24</td>
<td>+0.63</td>
<td>+0.12</td>
<td>+0.69</td>
</tr>
<tr>
<td>14. Cantley</td>
<td>16.36</td>
<td>-0.83</td>
<td>+0.62</td>
<td>+0.36</td>
<td>+0.92</td>
</tr>
<tr>
<td>Mean ...</td>
<td>11.53</td>
<td>+0.32</td>
<td>+0.07</td>
<td>+0.14</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Centre</th>
<th>Sugar Percentage</th>
<th>Mean Yield</th>
<th>Mean response to</th>
<th>Standard Error</th>
<th>Interactions</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>P</td>
<td>K</td>
<td>N x P</td>
<td>N x K</td>
</tr>
<tr>
<td>2. Ipswich</td>
<td>15.92</td>
<td>-0.84*</td>
<td>-0.38</td>
<td>+0.34</td>
<td>+0.320</td>
<td>-0.20</td>
</tr>
<tr>
<td>3. Colwick</td>
<td>15.10</td>
<td>-0.13</td>
<td>-0.12</td>
<td>+0.22</td>
<td>+0.322</td>
<td>+0.10</td>
</tr>
<tr>
<td>4. Newark</td>
<td>16.23</td>
<td>-0.38**</td>
<td>-0.22</td>
<td>+0.12</td>
<td>+0.119</td>
<td>-0.32</td>
</tr>
<tr>
<td>5. Felstead</td>
<td>16.72</td>
<td>-0.66**</td>
<td>+0.22</td>
<td>+0.14</td>
<td>+0.129</td>
<td>-0.35</td>
</tr>
<tr>
<td>6. Brigg</td>
<td>17.74</td>
<td>+0.36**</td>
<td>+0.09</td>
<td>+0.09</td>
<td>+0.096</td>
<td>+0.18</td>
</tr>
<tr>
<td>7. Poppleton</td>
<td>17.94</td>
<td>-0.15</td>
<td>-0.05</td>
<td>+0.25</td>
<td>+0.263</td>
<td>-0.24</td>
</tr>
<tr>
<td>8. Bardney</td>
<td>16.15</td>
<td>-0.62*</td>
<td>+0.01</td>
<td>+0.21</td>
<td>+0.286</td>
<td>+0.32</td>
</tr>
<tr>
<td>9. Allscott</td>
<td>15.93</td>
<td>-0.19</td>
<td>+0.02</td>
<td>+0.39**</td>
<td>+0.113</td>
<td>-0.08</td>
</tr>
<tr>
<td>10. Wissington</td>
<td>16.40</td>
<td>-0.62**</td>
<td>+0.12</td>
<td>+0.24</td>
<td>+0.166</td>
<td>-0.23</td>
</tr>
<tr>
<td>11. Peterborough</td>
<td>14.84</td>
<td>-0.15</td>
<td>+0.22</td>
<td>+0.20</td>
<td>+0.222</td>
<td>-0.53</td>
</tr>
<tr>
<td>12. King’s Lynn</td>
<td>16.69</td>
<td>+0.14</td>
<td>+0.63*</td>
<td>-0.24</td>
<td>+0.294</td>
<td>-0.48</td>
</tr>
<tr>
<td>13. Ely</td>
<td>15.89</td>
<td>-0.07</td>
<td>-0.04</td>
<td>-0.13</td>
<td>+0.189</td>
<td>-0.14</td>
</tr>
<tr>
<td>14. Cantley</td>
<td>15.70</td>
<td>-0.33</td>
<td>+0.07</td>
<td>+0.14</td>
<td>+0.212</td>
<td>-0.06</td>
</tr>
<tr>
<td>Mean ...</td>
<td>16.25</td>
<td>-0.34</td>
<td>+0.03</td>
<td>+0.15</td>
<td>-0.16</td>
<td>+0.09</td>
</tr>
</tbody>
</table>

- the first "damned stars"..?
Biological significance?

- 5 best 100m. times for Asafa Powell and Linford Christie
  - Powell is significantly faster
  - but would it matter if they were running for the bus?

Mann-Whitney U (Wilcoxon rank-sum) test

Variates: Powell, Christie.

Value of U: 0.0 (second sample has highest rank score).

Exact probability: 0.008 (under null hypothesis that Powell is equal to Christie).

Sample sizes: 5, 5.
Randomization

  - systematic designs introduce "a flagrant violation of the conditions upon which a valid estimate {of error} is possible" – se's "probably increase"
  - "The estimate of error is valid because, if we imagine a large number of different results obtained by different random arrangements, the ratio of the real to the estimated error, calculated afresh for each of these arrangements, will be actually distributed in the theoretical distribution by which the significance of the result is tested."
  - i.e. randomization distribution should approximate Normal distribution
  - demonstrated by Eden & Yates (1933, *J.Agric.Sci*) with uniformity data

- very controversial at the time

- Bailey (1983) Restricted randomization. *Biometrika*
  - how to rule out "unfortunate" randomizations
Replication

  - "It would be exceedingly inconvenient if every field trial had to be preceded by a succession of even ten uniformity trials; consequently since the only purpose of these trials is to provide an estimate of the standard error, means have been devised for obtaining such an estimate from the actual yields of the trial year. The method adopted is that of replication. ..."
  - also introduced the concept of blocking, and use of the Latin square

  - claimed that the significance test (by F distribution) in the Latin square was not unbiased
  - but these referred to his own null hypothesis, not Fisher's (whose F test is unbiased!)
Factorial design

• Fisher (1926) The arrangement of field experiments.

  *J. Min. Ag. G. Br.*

  • "No aphorism is more frequently repeated in connection with field trials, than that we should ask nature few questions or, ideally, one question, at a time. The writer is convinced that this view is mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed."

  • if there is no interaction, we have combined several one-at-a-time experiments

  • while, if there is an interaction, we are much wiser!

• also introduced confounding

  • "The comparisons to be sacrificed will be deliberately confounded with certain elements of the soil heterogeneity, and with them eliminated."
First design used in practice

- factorial in randomized blocks
- $2^3 + \text{control}$ (replicated 4 times)

---


<table>
<thead>
<tr>
<th>1M EARLY</th>
<th>1M LATE</th>
<th>1M EARLY</th>
<th>1M LATE</th>
<th>1M EARLY</th>
<th>1M LATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
</tr>
<tr>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
</tr>
<tr>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
</tr>
<tr>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
<td>2M EARLY</td>
<td>2M LATE</td>
</tr>
<tr>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
</tr>
<tr>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
</tr>
<tr>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
<td>1M EARLY</td>
<td>1M LATE</td>
</tr>
</tbody>
</table>

Fig. 1. A complex experiment with winter oats. (Reproduced from the *Journal of the Ministry of Agriculture* by permission of the Controller of H.M. Stationery Office.)
Analysis of variance

- distinguished between
  - plot error (24 d.f.) from within-block replicates of null control
  - block-treatment interaction ("differential responses")
Analysis of variance

Variate: Grain

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>s.s.</th>
<th>m.s.</th>
<th>v.r.</th>
<th>F pr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks stratum</td>
<td>7</td>
<td>2286.44</td>
<td>326.63</td>
<td>10.14</td>
<td></td>
</tr>
<tr>
<td>Blocks.Plots stratum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatments</td>
<td>8</td>
<td>387.01</td>
<td>48.38</td>
<td>1.50</td>
<td>0.209</td>
</tr>
<tr>
<td>Blocks.Treatments</td>
<td>56</td>
<td>2508.72</td>
<td>44.80</td>
<td>1.39</td>
<td>0.189</td>
</tr>
<tr>
<td>Residual</td>
<td>24</td>
<td>773.21</td>
<td>32.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>5955.39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- block-treatment interaction not significant
  - so can combine errors

..
**Analysis of variance**

Variate: Grain

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>s.s.</th>
<th>m.s.</th>
<th>v.r.</th>
<th>F pr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks stratum</td>
<td>7</td>
<td>2286.44</td>
<td>326.63</td>
<td>8.04</td>
<td></td>
</tr>
<tr>
<td>Blocks.Plots stratum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anytreat</td>
<td>1</td>
<td>191.75</td>
<td>191.75</td>
<td>4.72</td>
<td>0.033</td>
</tr>
<tr>
<td>Anytreat.Amount</td>
<td>1</td>
<td>24.84</td>
<td>24.84</td>
<td>0.61</td>
<td>0.437</td>
</tr>
<tr>
<td>Anytreat.Timing</td>
<td>1</td>
<td>76.84</td>
<td>76.84</td>
<td>1.89</td>
<td>0.173</td>
</tr>
<tr>
<td>Anytreat.Type</td>
<td>1</td>
<td>17.67</td>
<td>17.67</td>
<td>0.43</td>
<td>0.512</td>
</tr>
<tr>
<td>Anytreat.Amount.Timing</td>
<td>1</td>
<td>35.07</td>
<td>35.07</td>
<td>0.86</td>
<td>0.356</td>
</tr>
<tr>
<td>Anytreat.Amount.Type</td>
<td>1</td>
<td>7.48</td>
<td>7.48</td>
<td>0.18</td>
<td>0.669</td>
</tr>
<tr>
<td>Anytreat.Timing.Type</td>
<td>1</td>
<td>23.31</td>
<td>23.31</td>
<td>0.57</td>
<td>0.451</td>
</tr>
<tr>
<td>Residual</td>
<td>81</td>
<td>3292.00</td>
<td>40.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>5955.39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Significant effect of nitrogen
- but not of differences in Amount, Timing or Type
Later work - Frank Yates

- Yates (1933) The principles of orthogonality and confounding in designed experiments. *J. Agric. Sci*
  - (1933) "since it is logically impossible that an interaction should exist without a main effect, the significance of main effects should be tested strictly on the assumption that their interactions are negligible" (c.f. Nelder, 1977, A reformulation of linear models. *JRSS A*
  - tested interactions in non-orthogonal expts by "fitting constants"
  - confounding of main effects – split plots, strip plots, Latin square with additional treatment factors applied to rows, and to columns
  - confounding of interactions (to avoid blocks becoming too large)
  - partial and balanced confounding
  - balanced incomplete blocks, efficiency factor
  - Finlay & Wilkinson (63 *Aust.J.Agric.Res*) method to assess interactions
  - lattice designs, pseudo-factors, efficiency factors
Later work – general balance

- set the Fisher & Yates concepts of randomization, multiple sources of random error, balance, orthogonality, efficiency factors etc into a general theoretical framework

- references
Likelihood

  - used likelihood for estimation
  - maximum likelihood, consistency, sufficiency..
  - fiducial inference
- Ross (1987) *Maximum Likelihood Program*
- Ross (1990) *Nonlinear Estimation*. Springer
  - use of maximum likelihood at Rothamsted in practice
Non-Normal data

- Fisher (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. A.*
  - (nonlinear) design and analysis of dilution series experiments
  - maximum likelihood analysis of probit data
  - generalized linear models with several error terms, h-likelihood
Variance components

  - introduced concept of analysis of variance components
- Patterson & Thompson (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*
  - REML (residual/restricted maximum likelihood) algorithm
- Gilmour, Cullis & Verbyla (1997). Accounting for natural extraneous variation in analysis of field experiments. *JABES*
- Coombes, Payne, & Lisboa (2002) Comparison of nested simulated annealing and reactive tabu search for efficient experimental designs with correlated data. *COMPSTAT 2002*
REML – subnote

- also used in plant science
  - e.g. Baird, Johnstone, & Wilson (2004) Normalization of microarray data using a spatial mixed model analysis which includes splines. *BioInformatics*

- providing a link to Fisher's interests in genetics (not discussed here)
  - L.J. Savage (1976) *Annals of Statistics* "I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician"

..
Statistical computing

- Millionaire calculating machine
  - bought by Fisher at a cost >£200
  - used here by his successor Frank Yates (Head of Department 1933-1968)
  - Fisher: "most of my statistics has been learned on the machine"

..
- Elliott 401 computer installed at Rothamsted in 1954
  - Gower (1985) "first computer to be associated primarily with agricultural research and with statistics"
  - Yates (1955) "Having an electronic machine on the spot has made all the difference to developing its applications to research statistical problems"
Statistical computing – later work

GenStat for Windows
- developed by Rothamsted since 1968 (and now by VSN)
- see www.vsni.co.uk
Conclusion

• problems posed by the agricultural and biological research at Rothamsted have spurred many advances in statistics
  • tradition started by Fisher, and continued ever since
• Fisher recommended full statistical involvement
  • "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." Presidential Address to the First Indian Statistical Congress, 1938.
• also see e.g. Bailey & Payne: Experimental design: statistical research and its application. Institute of Arable Crops Research Report for 1989
  ..