# Chapter 5. Examples of Data on Permutations and Homogeneous Spaces

To fix ideas, as well as to make contact with reality, it is useful to have a collection of real data sets on hand.

A. PERMUTATION DATA.

(1) Large sets of rankings are sometimes generated in psychophysical experiments (rank these sounds for loudness), taste testing experiments (rank these 5 types of coffee ice cream), or surveys. To give an example, in 1972, the National Opinion Research Center included the following question in one of their surveys: Where do you want to live? Rank the following 3 options: in a big city; near a big city ($\leq$ 50 miles); far from a big city ($>$ 50 miles). The data from 1439 respondents was

| city | suburbs | country | # |
|---|---|---|---|
| 1 | 2 | 3 | 242 |
| 1 | 3 | 2 | 28 |
| 2 | 1 | 3 | 170 |
| 3 | 1 | 2 | 628 |
| 2 | 3 | 1 | 12 |
| 3 | 2 | 1 | 359 |

Let us briefly discuss this data. The modal rank is $\begin{smallmatrix}1 & 2 & 3\\ 3 & 1 & 2\end{smallmatrix}$ — people prefer the suburbs, then country, then city. This is born out by simple averages: 270 people ranked city first, 798 ranked suburb first, 371 ranked country first.

The 2 small counts lead to an interesting interpretation. Both violate the unfolding hypothesis of Coombs (1964). To spell this out a bit, suppose people's rankings are chosen in accordance with the ideal distance from the city, different people having different preferences. Thus, one chooses the rank one location and then "unfolds" around it. In this model $\left(\begin{smallmatrix}1 & 2 & 3\\ 1 & 3 & 2\end{smallmatrix}\right)$ is impossible since if one most prefers being in the city, one must prefer being close to the city to being far away. The number of permutations of the set $1, 2, \ldots, n$ consistent with unfolding is about $2^{n-1}$, so many arrangements are ruled out. Unfolding is a nice idea, but distance to the city might not determine things for someone who works in the suburbs and doesn't want to live where they work. If you ask people to rank order temperature for tea (hot, medium, cold), you don't expect the unfolding restriction to hold, but if you ask people to rank order sugar teaspoons (0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2) you do expect the data to be consistent with unfolding.

Further analysis of the distance to cities data is in Chapter 8. Duncan and Brody (1982) discuss these data in some detail.

Ranked data often comes with other variables — rankings for men and women, or by income being examples. In the data on distance to cities, the actual dwelling place of the respondent is available. Methods for dealing with covariates are developed in Chapter 9.

It is worth pointing to a common problem not represented in the cities data. Because n! grows so rapidly, one can have a fairly large data set of rankings and still only have a small proportion of the possible orders represented. For example, I am considering a data set in which 129 black students and 98 white students were asked to rank "score, instrument, solo, benediction, suite" from the least related to "song" to the most strongly related to "song." Here, there cannot be very many repeats in each ranking. In another data set, quoted in Feigin and Cohen (1978), 148 people ranked 10 occupations for desirability. Clearly, the ratio of the sample size to n! has a limiting effect on what kind of models can be fit to the data.

(2) Pairs of permutations often arise as in "rank order the class on the midterm and final." Similarly, small sets of rankings arise as in a panel of judges ranking a set of contestants. A large collection of examples appears in Chapter 7A.

(3) *The Draft Lottery.* In 1970, a single "random" permutation in $S_{365}$ was chosen. This permutation was used to fix the order of induction into the army. The actual permutation is shown in Table 1. For discussion of this data set, see the article by S. E. Fienberg (1971).

As Fienberg reports, it was widely claimed that the permutation tended to have lower order months Jan., Feb., ... having higher numbers. The Spearman rank correlation coefficient is -.226, significant at the .001 level. Figure 2, based on Figure 1, shows the average lottery number by month. The evidence seems strong until we reflect on the problems of pattern finding in a single data source after agressive data analysis.

Further analysis of this data is given in example 1 of Chapter 7A.

## B. Partially ranked data.

There are numerous examples in which people rank a long list only partially. For example, people might be asked to rank their favorite 10 out of 40 movies, a typical ranking yielding $(a_1, a_2, \ldots, a_{10})$ with $a_1$ the name of the movie ranked first, etc. Alternatively people might be asked to choose a committee of 10 out of 40, not ranking within. Then a typical selection yields the set $\{a_1, a_2, \ldots, a_{10}\}$.

In each case the symmetric group $S_{40}$ acts transitively on the partial rankings which may thus be represented as homogeneous spaces for $S_{40}$ (see Chapter 3-F for definitions). For ranked 10 out of 40 the homogeneous space is $S_{40}/S_{30}$. For unranked 10 out of 40, the homogeneous space is $S_{40}/S_{10} \times S_{30}$.

Here are some real examples of such data.

*Example 1. American Psychological Association data.* The American Psychological Association is a large professional group (about 50,000 members). To vote for a president, members rank order five candidates. A winner is chosen by the *Hare system*: Look at the first place votes for all five candidates. If there is no majority candidate ($\geq 50\%$) delete the candidate with the fewest first place votes. Ballots

Figure 1
The 1970 Random Selection Sequence by Month and Day

| Day | Jan. | Feb. | Mar. | Apr. | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 305 | 086 | 108 | 032 | 330 | 249 | 093 | 111 | 225 | 359 | 019 | 129 |
| 2 | 159 | 144 | 029 | 271 | 298 | 228 | 350 | 045 | 161 | 125 | 034 | 328 |
| 3 | 251 | 297 | 267 | 083 | 040 | 301 | 115 | 261 | 049 | 244 | 348 | 157 |
| 4 | 215 | 210 | 225 | 081 | 276 | 020 | 279 | 145 | 232 | 202 | 266 | 165 |
| 5 | 101 | 214 | 293 | 269 | 364 | 028 | 188 | 054 | 082 | 024 | 310 | 056 |
| 6 | 224 | 347 | 139 | 253 | 155 | 110 | 327 | 114 | 006 | 087 | 076 | 010 |
| 7 | 306 | 091 | 122 | 147 | 035 | 085 | 050 | 168 | 008 | 234 | 051 | 012 |
| 8 | 199 | 181 | 213 | 312 | 321 | 366 | 013 | 048 | 184 | 283 | 097 | 105 |
| 9 | 194 | 338 | 317 | 219 | 197 | 335 | 277 | 106 | 263 | 342 | 080 | 043 |
| 10 | 325 | 216 | 323 | 218 | 065 | 206 | 284 | 021 | 071 | 220 | 282 | 041 |
| 11 | 329 | 150 | 136 | 014 | 037 | 134 | 248 | 324 | 158 | 237 | 046 | 039 |
| 12 | 221 | 068 | 300 | 346 | 133 | 272 | 015 | 142 | 242 | 072 | 066 | 314 |
| 13 | 318 | 152 | 259 | 124 | 295 | 069 | 042 | 307 | 175 | 138 | 126 | 163 |
| 14 | 238 | 004 | 254 | 231 | 178 | 356 | 331 | 198 | 001 | 294 | 127 | 026 |
| 15 | 017 | 039 | 169 | 273 | 130 | 180 | 322 | 102 | 113 | 171 | 131 | 320 |
| 16 | 121 | 212 | 166 | 148 | 055 | 274 | 120 | 044 | 207 | 254 | 107 | 096 |
| 17 | 235 | 189 | 033 | 260 | 112 | 073 | 058 | 154 | 255 | 288 | 143 | 304 |
| 18 | 140 | 292 | 332 | 090 | 278 | 341 | 190 | 141 | 246 | 005 | 146 | 128 |
| 19 | 058 | 025 | 200 | 236 | 075 | 104 | 227 | 311 | 177 | 241 | 203 | 240 |
| 20 | 280 | 302 | 239 | 346 | 123 | 360 | 187 | 344 | 063 | 192 | 185 | 135 |
| 21 | 186 | 363 | 334 | 062 | 250 | 060 | 027 | 291 | 204 | 243 | 156 | 070 |
| 22 | 337 | 290 | 265 | 316 | 326 | 247 | 153 | 339 | 160 | 117 | 009 | 053 |
| 23 | 118 | 057 | 256 | 252 | 319 | 109 | 172 | 116 | 119 | 201 | 182 | 162 |
| 24 | 059 | 236 | 258 | 002 | 031 | 358 | 023 | 036 | 195 | 196 | 230 | 095 |
| 25 | 052 | 179 | 343 | 351 | 361 | 137 | 067 | 286 | 149 | 176 | 132 | 084 |
| 26 | 092 | 365 | 170 | 340 | 357 | 022 | 303 | 245 | 018 | 007 | 309 | 173 |
| 27 | 355 | 205 | 268 | 074 | 296 | 064 | 289 | 352 | 233 | 264 | 047 | 078 |
| 28 | 077 | 299 | 223 | 262 | 308 | 222 | 088 | 167 | 257 | 094 | 281 | 123 |
| 29 | 349 | 285 | 362 | 191 | 226 | 353 | 270 | 061 | 151 | 229 | 099 | 016 |
| 30 | 164 | | 217 | 208 | 108 | 209 | 287 | 333 | 315 | 038 | 174 | 003 |
| 31 | 211 | | 030 | | 313 | | 193 | 011 | | 079 | | 100 |

Figure 2

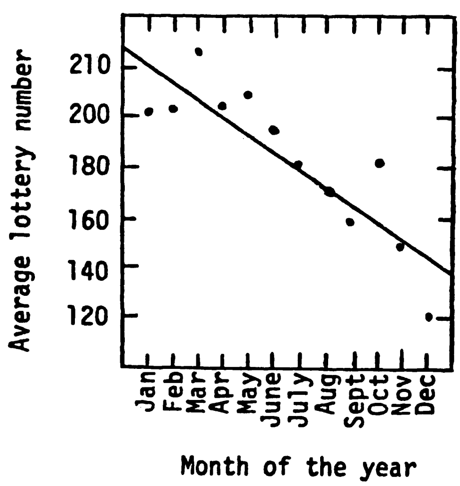


Fig. 2. Average lottery numbers by month. The line is the least squares regression line, treating the months as being equally spaced.

with this candidate are relabelled to have the remaining candidates in the same relative order. The procedure is now continued with the four remaining candidates. Fishburn (1973), Doran (1979), or Brams and Fishburn (1983) discuss the system and relevant literature.

A considerable number of voters do not rank all five candidates. For example, in the year being considered the number of voters ranking $q$ of the candidates was

| $q$ | # | |
|---|---|---|
| 1 | 5141 | |
| 2 | 2462 | 9711 |
| 3 | 2108 | |
| 5 | 5738 | |
| | 15,449 | |

Thus there were 5,738 complete rankings, but 5,141 only voted for their first choice. In all, more than half of the ballots were incomplete. It is assumed that people who rank 4 candidates meant to rank the 5th candidate last.

It is natural to inquire whether the partially ranked ballots are different from the restriction of the complete ballots (or vary with $q$). Such considerations should play a role in deciding on a final voting rule, and on deciding on ballot design and election publicity in following years.

Table 1 gives the complete data. The data are arranged as (rank, #) where rank is a five-digit number, whose ith digit represents the rank given to candidate $i$ (a zero or blank means that this is a partial ranking, in which candidate $i$ has not been ranked). For example, the first entry (1, 1022) indicates that candidate 5 was ranked first by 1022 people who didn't rank anyone else. The second entry (10, 1145) indicates that candidate 4 was ranked first by 1145 people (who didn't rank anyone else). The first 5 entries give the totals for singly ranked items. The next 20 entries give totals for people ranking 2 of the 5 candidates. For example 143 people ranked candidate 5 first and candidate 4 second (and didn't rank anyone else). These data are analyzed by Diaconis (1989).

*Example 2. k sets of an n set.* If people are asked to choose their favorite $k$ of $n$, *without* ranking within (as in choosing a committee or set of invitees to a meeting), then the relevant homogeneous space is $S_n/S_k \times S_{n-k}$, where $S_k \times S_{n-k}$ is the subgroup of $S_n$ allowing arbitrary permutations among $\{1, \ldots, k\}$ and among $\{k+1, \ldots, n\}$. Approval voting, recommended by Brams and Fishburn (1983) yields such data.

Here is an example where large amounts of such data occur. The State of California has a state lottery game called 6/49 or Lotto. To play, you select a 6 set from $\{1, 2, \ldots, 49\}$. Then, 6 of 49 numbered balls are chosen at random. The grand prize is divided between the people choosing this subset.

There are about 14 million subsets, and 11 million players per week in this game at present. Of course, people do not choose subsets at random — they play favorate combinations. One can get a distinct advantage in this game by avoiding popular numbers and subsets. After all, if you are the only person on the subset you don't have to split with anyone. This can actually overcome the "house take"

Table 1
American Psychological Association Election Data

| Partial Ranking | # of Votes Cast of This Type | Partial Ranking | # of Votes Cast of This Type | Partial Ranking | # of Votes Cast of This Type | Partial Ranking | # of Votes Cast of This Type |
|---|---|---|---|---|---|---|---|
| 1 | 1022 | 23100 | 83 | 45213 | 24 | 24135 | 96 |
| 10 | 1145 | 20103 | 74 | 45132 | 38 | 23541 | 45 |
| 100 | 1198 | 132 | 19 | 45123 | 30 | 23514 | 52 |
| 1000 | 881 | 123 | 15 | 43521 | 91 | 23451 | 53 |
| 10000 | 895 | 2103 | 16 | 43512 | 84 | 23415 | 52 |
| 21 | 143 | 1302 | 15 | 43251 | 30 | 23154 | 186 |
| 12 | 196 | 1032 | 45 | 43215 | 35 | 23145 | 172 |
| 201 | 64 | 1320 | 17 | 43152 | 38 | 21543 | 36 |
| 210 | 48 | 1203 | 8 | 43125 | 35 | 21534 | 42 |
| 102 | 93 | 31002 | 38 | 42531 | 58 | 21453 | 24 |
| 120 | 56 | 31020 | 45 | 42513 | 66 | 21435 | 26 |
| 2001 | 70 | 31200 | 32 | 42351 | 24 | 21354 | 30 |
| 2010 | 114 | 21003 | 17 | 42315 | 51 | 21345 | 40 |
| 2100 | 89 | 21030 | 31 | 42153 | 52 | 15432 | 40 |
| 1002 | 80 | 1023 | 55 | 42135 | 40 | 15423 | 35 |
| 1020 | 87 | 1230 | 9 | 41532 | 50 | 15342 | 36 |
| 1200 | 51 | 21300 | 31 | 41523 | 45 | 15324 | 17 |
| 20001 | 117 | 10032 | 35 | 41352 | 31 | 15243 | 70 |
| 20010 | 104 | 10203 | 49 | 41325 | 23 | 15234 | 50 |
| 20100 | 547 | 10302 | 41 | 41253 | 22 | 14532 | 52 |
| 21000 | 72 | 10320 | 21 | 41235 | 16 | 14523 | 48 |
| 10002 | 72 | 13002 | 31 | 35421 | 71 | 14352 | 51 |
| 10020 | 74 | 13020 | 22 | 35412 | 61 | 14325 | 24 |
| 10200 | 302 | 13200 | 79 | 35241 | 41 | 14253 | 70 |
| 12000 | 83 | 10023 | 44 | 35214 | 27 | 14235 | 45 |
| 30021 | 75 | 10230 | 30 | 35142 | 45 | 13542 | 35 |
| 30201 | 32 | 12003 | 26 | 35124 | 36 | 13524 | 28 |
| 32001 | 41 | 12030 | 19 | 34521 | 107 | 13452 | 37 |
| 20031 | 62 | 12300 | 27 | 34512 | 133 | 13425 | 35 |
| 20301 | 37 | 54321 | 29 | 34251 | 62 | 13254 | 95 |
| 23001 | 35 | 54312 | 67 | 34215 | 28 | 13245 | 102 |
| 3201 | 15 | 54231 | 37 | 34152 | 87 | 12543 | 34 |
| 2301 | 14 | 54213 | 24 | 34125 | 35 | 12534 | 35 |
| 3021 | 59 | 54132 | 43 | 32541 | 41 | 12453 | 29 |
| 2031 | 50 | 54123 | 28 | 32514 | 64 | 12435 | 27 |
| 321 | 20 | 53421 | 57 | 32451 | 34 | 12354 | 28 |
| 231 | 17 | 53412 | 49 | 32415 | 75 | 12345 | 30 |
| 30012 | 90 | 53241 | 22 | 32154 | 82 | | |
| 30210 | 13 | 53214 | 22 | 32145 | 74 | | |
| 32010 | 51 | 53142 | 34 | 31542 | 30 | | |
| 20013 | 46 | 53124 | 26 | 31524 | 34 | | |
| 20310 | 15 | 52431 | 54 | 31452 | 40 | | |
| 23010 | 28 | 52413 | 44 | 31425 | 42 | | |
| 3012 | 62 | 52341 | 26 | 31254 | 30 | | |
| 3210 | 18 | 52314 | 24 | 31245 | 34 | | |
| 2310 | 21 | 52143 | 35 | 25431 | 35 | | |
| 2013 | 54 | 52134 | 50 | 25413 | 34 | | |
| 312 | 46 | 51432 | 50 | 25341 | 40 | | |
| 213 | 16 | 51423 | 46 | 25314 | 21 | | |
| 2130 | 17 | 51342 | 25 | 25143 | 106 | | |
| 3120 | 26 | 51324 | 19 | 25134 | 79 | | |
| 3102 | 16 | 51243 | 11 | 24531 | 63 | | |
| 30102 | 47 | 51234 | 29 | 24513 | 53 | | |
| 32100 | 57 | 45321 | 31 | 24351 | 44 | | |
| 30120 | 15 | 45312 | 54 | 24315 | 28 | | |
| 20130 | 39 | 45231 | 34 | 24153 | 162 | | |

and yield a favorable game. Chernoff (1981) gives details for the Massachusetts lottery. In any case, the data must be analyzed.

While it is not possible to present such data here, the following smaller example shows that interesting analyses are possible.

There are various gadgets sold to generate a six-element subset of $\{1, 2, \ldots, 49\}$. These are used to help players pick combinations for the California state Lotto game.

One such gadget is pictured in Figure 1. There are 49 numbered holes and six balls enclosed by a plastic cover. One shakes the balls around and uses the six set determined by their final resting place.

Figure 1.

| PICK 6 | | | LOTTO | | | | & WIN | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| o | o | o | o | o | o | o | o | o |
| 10 | 11 | 12 | 13 | | 14 | 15 | 16 | 17 |
| o | o | o | o | | o | o | o | o |
| 18 | 19 | 20 | 21 | | 22 | 23 | 24 | 25 |
| o | o | o | o | | o | o | o | o |
| 26 | 27 | 28 | 29 | | 30 | 31 | 32 | 33 |
| o | o | o | o | | o | o | o | o |
| 34 | 35 | 36 | 37 | | 38 | 39 | 40 | 41 |
| o | o | o | o | | o | o | o | o |
| 42 | 43 | 44 | 45 | | 46 | 47 | 48 | 49 |
| o | o | o | o | | o | o | o | o |

This gadget seems at first like other classical devices to generate random outcomes: if vigorously shaken, it should lead to random results. Further thought suggests that the outer, or border numbers might be favored over the inner numbers.

To test this, 100 trials were performed. The gadget was vigorously shaken and set down on a flat surface. The results are given in Table 2.

Following each six set is $X$ — the number of balls falling on the outer perimeter in that 6-set. For example, the first 6-set $\{10, 11, 13, 25, 36, 42\}$ had 3 outside numbers — 10, 25, 42 — so $X = 3$. There are 25 outside numbers out of 49.

Table 2

100 6−sets of $\{1,2,\ldots,49\}$

| | | | |
|---|---|---|---|
| 10,11,13,25,36,42/3 | 5,10,21,26,42,46/5 | 4,17,18,22,32,41/4 | 1,17,22,25,29,31/3 |
| 25,27,34,39,45,46/4 | 16,23,37,41,43,45/3 | 6, 9,10,12,16,32/3 | 2,13,23,24,26,30/2 |
| 3, 5,18,20,33,39/4 | 8,10,13,34,43,49/5 | 2, 5,17,19,36,40/3 | 2, 6,15,18,32,37/3 |
| 3,10,23,26,45,49/5 | 2,10,11,12,13,15/2 | 2, 6,10,25,33,38/5 | 2,14,15,17,18,35/3 |
| 3, 7,15,19,26,34/4 | 10,13,15,22,26,43/3 | 3,17,29,40,41,45/4 | 4,10,20,31,32,37/2 |
| 4,15,32,33,36,49/3 | 15,19,22,30,32,39/0 | 4, 7,11,23,35,36/2 | 7,13,17,27,31,44/2 |
| 10,11,23,33,43,46/4 | 2,15,22,25,29,48/3 | 1,18,31,33,34,46/5 | 19,22,28,32,42,44/2 |
| 1, 6, 7,18,26,34/6 | 6, 7,10,11,17,31/3 | 11,13,15,28,34,39/1 | 7,13,19,33,47,48/4 |
| 6,11,15,19,26,46/3 | 6,17,24,29,42,43/4 | 4, 7,15,18,31,33/4 | 1, 2, 4,15,19,40/3 |
| 1,11,15,18,26,29/3 | 2, 9,21,36,43,45/4 | 1, 3,12,15,20,41/3 | 2, 5,25,26,30,39/4 |
| 10,11,19,31,36,42/2 | 7,12,18,35,42,44/4 | 4, 9,12,22,39,41/3 | |
| 6,15,25,27,42,47/4 | 5,16,18,33,36,39/3 | 3,10,12,28,34,39/3 | |
| 17,18,33,36,43,46/5 | 2, 6, 7,11,31,47/4 | 2, 7,12,27,34,35/3 | |
| 1, 2,32,36,43,48/4 | 18,22,28,36,42,47/3 | 1, 4, 7,12,20,43/4 | |
| 16,20,30,35,45,46/2 | 4,18,29,35,39,46/3 | 5, 7,14,16,18,31/3 | |
| 16,20,26,37,42,49/3 | 3, 6,16,25,29,42/4 | 6,23,28,34,36,40/2 | |
| 3,18,27,30,42,43/4 | 1,28,31,37,42,43/3 | 2, 5, 9,15,23,27/3 | |
| 9,10,27,42,43,45/5 | 1,18,23,27,42,43/4 | 2, 3,19,34,39,44/4 | |
| 6,23,32,39,42,46/3 | 4, 5, 7, 8,40,42/5 | 6,12,14,16,23,39/1 | |
| 5,19,36,39,42,44/3 | 6, 7, 9,12,39,49/4 | 2,12,15,26,38,43/3 | |
| 7,18,20,29,35,43/3 | 12,13,18,19,22,36/1 | 5, 7,12,17,29,35/3 | |
| 12,14,23,29,41,48/2 | 4, 7, 8,10,33,49/6 | 4, 9,16,23,27,42/3 | |
| 4, 6,17,20,33,48/5 | 7, 9,31,32,41,46/4 | 2,13,15,20,21,48/2 | |
| 4,18,27,30,43,49/4 | 9,12,14,37,46,48/3 | 1, 5,34,42,44,46/6 | |
| 6,10,18,30,35,45/4 | 7, 9,16,29,41,46/4 | 15,16,17,24,27,30/1 | |
| 26,27,38,42,43,44/4 | 14,19,21,28,33,42/2 | 8,15,18,21,30,39/2 | |
| 6, 8,19,38,43,49/4 | 6,14,15,17,31,49/3 | 6,15,21,23,32,47/2 | |
| 1,20,25,42,43,49/5 | 8,33,35,41,45,47/5 | 5, 7, 8,19,23,49/4 | |
| 4,34,27,39,43,46/4 | 2, 8,25,29,42,47/5 | 7, 8, 9,14,20,22/3 | |
| 3, 4,11,33,46,49/5 | 8,11,24,25,37,48/3 | 10,15,29,34,46,49/4 | |

If the six sets were chosen at random, $X$ would have a hypergeometric distribution $H\{X=j\} = \frac{\binom{25}{j}\binom{24}{6-j}}{\binom{49}{6}}$. These numbers are given in Table 3 which also shows the empirical counts from Table 2.

Table 3

Hypergeometric and Empirical Probabilities for $X$.

| j | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $H\{X=j\}$ | .013 | .091 | .250 | .333 | .228 | .016 | .010 |
| Empirical | .01 | .04 | .14 | .35 | .30 | .13 | .03 |

The differences are *not* overwhelming visually. They do show up in two straightforward tests.

A first test was based on $p = H\{X = 4,5,6\} = .353$ versus the empirical ratio $\hat{p} = .46$. Then $(\hat{p}-p)/\sqrt{p(1-p)/100} = 2.23$. This difference, more than two standard deviations, is convincing evidence against uniformity.

Colin Mallows suggested using the average, $\overline{X}$, as a statistic. Under the null distribution, $E(\overline{X}) \doteq 3.06$, $SD(\overline{X}) \doteq 0.116$. The observed $\overline{X}$ is 3.40. This yields a standardized ($z$ value) of 2.92.

**Remarks.**

1) As is well known, the omnibus chi-square test is to be avoided for these kinds of problems. Because it tries to test for all possible departures from uniformity, chi-square only works well for large deviations or sample sizes. Interestingly, here it *fails* to reject the null (10.23 on six degrees of freedom

with all 7 categories or 9.01 on five degrees of freedom with the first and last categories combined).

2) Other questions can be asked of these data. To begin with, the central numbers

$$\begin{array}{cccc} 20, & 21, & 22, & 23 \\ 28, & 29, & 30, & 31 \end{array}$$

presumably occur less often. More generally, a test that looks at all numbers, but takes into account the distance from the edge, could be constructed. A preliminary graphical analysis was *not* instructive.

Interesting questions arise about the corners and about individual numbers. With more data, some second order questions can be entertained.

3) It seems clear that this style of randomization mechanism is badly flawed. Possible physical explanations can be entertained to explain these flaws. The balls lose most of their energy on impact with the sides, and then "trickle back" to the edge. A slight tilt draws the balls toward an edge.

4) One practical application of this kind of testing problem comes in the actual lottery. A quick test to detect marked departures is needed for a pre-game screening (someone might have switched for loaded balls during the night).

*Example 3. Q sort data.* The General Social Survey lists thirteen qualities a child could possess. From this list, respondents are asked to choose the most desirable quality, the two next most desirable qualities, the least desirable quality and the next two least desirable qualities. In an obvious way, this is data on $S_{13}/S_1 \times S_2 \times S_7 \times S_2 \times S_1$. More generally, if $\lambda$ is a partition of $n$, so $\lambda = (\lambda_1, \ldots, \lambda_m)$ with $\lambda_1 + \ldots + \lambda_m = n$, one can consider data of the form: choose the first $\lambda_1$ objects (but do not order between), choose the next $\lambda_2$ objects, etc., finishing with $\lambda_m$ objects ranked last. Such a scheme is called $Q$ sort data in psychology experiments. It is not unusual to ask for a list of 100 items to be ranked for its degree of concordance or similarity with a fixed object. For example, the object might be a person (spouse, national leader) and the items might be descriptive levels of aggression. Suppose 9 categories of similarity are used, ranging from 1 - "most uncharacteristic," through 5 "neither characteristic nor uncharacteristic," up to 9 - "most characteristic." To aid in different rates, a forced distribution is often imposed. For $n = 100$, the numbers permitted in each category are often chosen from binomial considerations as 5, 8, 12, 16, 18, 16, 12, 8, 5. A novel application and references to the older literature may be found in L. E. Moses et al (1967). For more recent discussion see Heavlin (1980).

*Example 4. Other actions of $S_n$.* The symmetric group acts on many other combinatorial objects, such as the set of partitions or labelled binary trees. It follows that there is a wide variety of objects to which the analysis of this and succeeding chapters may be applied.

## C. THE $d$-SPHERE $S^d$.

Sometimes data are collected on the circle – which way do birds leave their nests. Data are also collected on the sphere – for example, in investigating the theory of continental drift, geologists looked at magnetization direction of rock samples on two "sides" of a purported boundary. Roughly, small pieces of certain kinds of rocks have a given magnetic orientation giving points on the sphere in $\mathbb{R}^3$. This leads to two-sample and other data analytic problems. Such considerations led Fisher (1953) to invent his famous family of distributions on the sphere.

Here is an example of data on higher dimensional spheres: consider testing whether measurement errors are normal. Samples of size $p$ are available from a variety of different sources. Say sample $i$ is normal with parameters $\mu_i$, $\sigma_i^2$:

$$\begin{aligned}&(X_{11},\ldots,X_{1p}) \text{ i.i.d. } n(\mu_1,\sigma_1^2)\\&(X_{21},\ldots,X_{2p}) \text{ i.i.d. } n(\mu_2,\sigma_2^2)\\&(X_{n1},\ldots,X_{np}) \text{ i.i.d. } n(\mu_p,\sigma_p^2).\end{aligned}$$

Think of $p$ small (say 10) and $n$ large (say 50). All samples are assumed independent. Let $\overline{X}_i$ and $S_i$ be the $i$th sample mean and standard deviation.

$$Y_i = \left(\frac{X_{i1}-\overline{X}_i}{S_i},\ldots,\frac{X_{ip}-\overline{X}_i}{S_i}\right).$$

The spherical symmetry of the normal distribution implies that $Y_i$ are randomly distributed over a $p-2$ dimensional sphere. Standard tests for uniformity thus provide tests for normality.

The group of $n\times n$ orthogonal matrices $O(n)$ acts transitively on the $n$ sphere. The subgroup fixing a point (say the north pole $(1,0,\ldots,0)$) is clearly $O(n-1)$. Thus the sphere can be thought of as $O(n)/O(n-1)$ and the rich tools of harmonic analysis become available.

Further introductory discussion is in Chapter 9B. Mardia (1972) and Watson (1983) give motivated, extensive treatments of data on the sphere.

## D. Other groups.

Many other groups occur. For example binary test results (e.g. right/wrong on the ith question $1 \leq i \leq k$) lead to data on $Z_2^k$. Here, for $x \in Z_2^k$, $f(x)$ is the number of people answering with pattern $x$. In *panel studies* a subject is followed over time. For example, 5,000 people may be followed for a year, each month a one or zero is recorded as the person is employed or not. This leads to data on $Z_2^{12}$.

There is a curious data set for $Z_{365}\times Z_{365}$ connected to the birthday-deathday question. Some researchers claim famous people tend to die close to the date of their birth. See Diaconis (1985) for a review of this literature.

Data on yet other groups arises in testing Monte Carlo algorithms for generating from the uniform distribution. Such group valued random variables are useful in doing integrals over groups. Testing a generator leads to a sample on the group in question. I have looked at data for the orthogonal and unitary groups in this regard.

It seems inevitable that data on other groups and homogeneous spaces will arise naturally in applications. One final example: with many scatterplots, one has many covariance matrices. The set of positive definite $2 \times 2$ matrices is usefully represented as $GL_2/O_2$. Several other examples are given in the following chapters.

E. STATISTICS ON GROUPS.

The examples described above suggest a wealth of statistical problems. In classical language, there is

- Testing for uniformity (is the sample really random?)
- Two sample tests (is there a difference between men and women's rankings?)
- Assessing association (is husband's ranking close to wife's?)
- Model building (can this huge list of data be summarized by a few parameters?)
- Model testing

More inclusively, there is the general problem of data analysis: how to make sense of this type of data; how to discover structure and find patterns.

The next four chapters offer three different approaches to these problems. Chapter 6 develops measures of distance on groups and homogeneous spaces. These are used to carry all sorts of familiar procedures into group valued examples.

Chapter 8 develops an analog of the spectral analysis of time series for group valued data. This is explored in the examples of partially ranked data. These examples make full use of the representation theory of the symmetric group. Chapter 7 is devoted to a self-contained development of this theory.

Chapter 9 uses representation theory to develop a natural family of models. In familiar cases, these reduce to models introduced by applied workers. The theory shows how to go further, and gives a unified development for all groups at once.

Of course, there is no substitute for trying things out in real examples, where special knowledge and insight can be brought to bear. There has not been much Bayesian work on these problems that I know of. The problems of developing natural prior distributions with respect to invariance seem fascinating. Consonni and Dawid (1985) or Fligner and Verducci (1988) offer steps in this direction.