Research Report

PARTICIPATION RATES AND MAXIMAL PERFORMANCE: A Log-Linear Explanation for Group Differences, Such as Russian and Male Dominance in Chess

Neil Charness[1] and Yigal Gerchak[2]

[1]Department of Psychology, Florida State University, and [2]Department of Management Sciences, University of Waterloo

Abstract—*Can the superiority of some countries and groups at certain activities be explained solely by the relative sizes of the participating populations? We focus on the expected highest achievement,* **max**, *as a function of the participating group's size For several relevant statistical distributions,* max *can be shown to be approximately log-linear in sample size, with a slope of about 0 7* SD *units We use this relation* (max *is log-linear ~0 7 MILL7) to examine differences in performance in chess by men and women and by different countries The expected differences under MILL7 are very close to the observed differences We also examine the implications of MILL7 for the interpretation of other group differences and discuss its limitations*

The Chess Olympiad in Moscow in December 1994 was won, predictably, by the Russian chess team (Russia I) Russia's second-string team took third place (The U S team, bolstered by Russian emigrants, finished in sixth place behind Bosnia, England, and Bulgaria) With the exception of Bobby Fischer's brief reign, why have Soviet and Russian men dominated the world chess championship since the middle 1940s? Why, too, do men consistently beat women in chess, so much so that the world chess body has held separate competitions and championships for men and women? (The best woman player in the world is typically about a full standard deviation below the best man)

Rather than postulate cultural or innate explanations of superiority, we want to consider a simple yet important artifact the effect of participation rates on the expected achievement levels for *max*, the score of the top performer It is important to stress at the outset that our working assumption is that even if there is no difference (same mean and standard deviation of individual achievement) in underlying ability between the identified groups (which we are not claiming to be the case), large differences in maximal performance can arise solely from differing participation rates in the target activity This result implies that before group differences in performance can be assessed properly, group sizes should be taken into account

It is possible to quantify the relationship between sample size and extreme best performances reasonably precisely to yield a function that predicts differences in performance as a function of differences in sample size We have dubbed this relationship the "maximum/minimum is log-linear, with a slope ~0 7" (MILL7) function

Address correspondence to Neil Charness, Department of Psychology, Florida State University, Tallahassee, FL 32306-1051, e-mail charness@psy fsu edu

THE MODEL MILL7

Let X_i be the random achievement of individual i, $i = 1, \ldots, n$ The random variables $X_1, \ldots, X_n$ are assumed independent and identically distributed with common cumulative distribution function F, and probability density function f Let $Y = max\ \{X_1, \ldots, X_n\}$ We are interested in the rate at which expectation $E(Y) \equiv E_n$ grows with n, for large values of n In general,

$$E_n = n \int_{-\infty}^{\infty} x f(x) \{F(x)\}^{n-1} dx$$

Although E_n is always increasing in n, the rate of this increase will depend on the type of distribution X has For several commonly used distributions (exponential, logistic, normal), it is known that for large n, E_n is approximately linear in log n (e g , Arnold, Balakrishnan, & Nagaraja, 1992) In particular, for a normal distribution with mean μ and variance σ^2, we have approximately[1] (Arnold et al , 1992, Joshi & Balakrishnan, 1981, Tippet, 1925)

$$E_n = \mu + (1.25 + 0.66 \log_{10} n)\sigma \tag{1}$$

For the highest score, *max*, if n is increased 10-fold, E_n grows by approximately 0 66σ Note that this implies that the larger the variability in individuals' achievement, the larger the effect of population size

Also, if n is large, then for any small k, the expected achievement of the kth best of normally distributed variables is also approximately linear in log n, although the coefficient of log n will depend on k (Arnold et al , 1992) Figure 1, produced from a subset of values given in the appendix of Harter (1961), shows that the slope tends to increase with increases in rank

This relationship gives us a metric for assessing departures from our assumptions of sampling from a population with the same mean and standard deviation It enables us to predict the results of increases in participation rates when samples come from the same underlying normal (also exponential or logistic) distribution When the distribution is normal or logistic, the same relationship holds for the minimum, *min*, the score of the

[1] Unlike for some other distributions, the approximate log-linearity of E_n for the normal distribution is not readily apparent from casual reading of the order-statistics literature Nevertheless, this property is at least implicit in some of the more elaborate approximations provided there, and can be ascertained by regressing the exact values given there against log n, or verified by Monte Carlo simulation

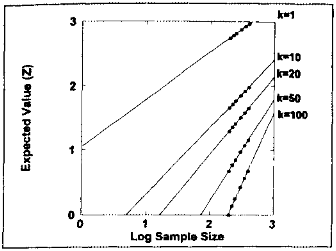

Fig 1 Expected value (Z score) for a standard normal distribution as a function of $\log_{10}$ sample size (N) for selected values of N and k (the rank order within the distribution) Data were obtained from Harter (1961) Approximate slopes are 0 738 for $k = 1$, 1 04 for $k = 10$, 1 20 for $k = 20$, 1 57 for $k = 50$, and 2 22 for $k = 100$

worst performer Logan (1988) discussed the log-linear relationship for *min* in the context of minimum reaction time and instance theory

CHESS RATINGS

Chess ratings are acquired when players compete in sanctioned tournaments under standardized conditions Rating points are lost or won based on winning or losing (or drawing) chess games against opponents who have ratings The amount won or lost depends on the difference in rating between the opponents The assumption made is that chess ability is approximately normally distributed, with the probability of winning dependent on the difference in ability, which is also assumed to be distributed roughly normally The computations are based on a logistic function (Elo, 1986, p 147)

Chess ratings form an interval scale, with the standard deviation set to 200 rating points Elo (1986) provided the distribution for established players (those with at least 20 games in rated tournaments) belonging to the United States Chess Federation (USCF) in 1977, a time at which, he stated, there was parity between these ratings in the USCF and those in the Fédération Internationale des Echecs (FIDE), the world chess fed-

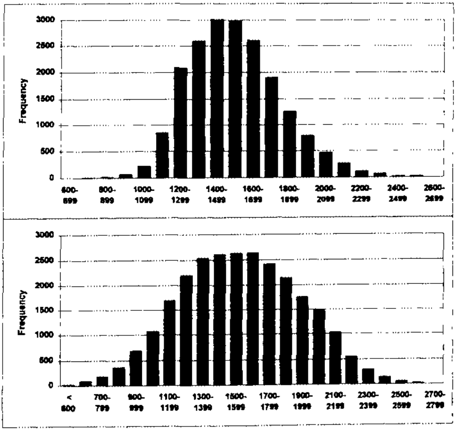

Fig 2 Histogram for the chess ratings of active players in the United States Chess Federation in 1977 (N = 19,405, top) and 1990 (N = 26,655, bottom)

eration As the top of Figure 2 shows, the distribution is somewhat skewed to the right, with more strong players than weak ones Elo (1986) reported that the mean was 1547, *SD* = 252, and that a chi-square test rejected the assumption that the distribution was normal There were too many strong and not enough weak players Selective attrition from tournament play could explain this departure from normality

As a comparison, the bottom of Figure 2 gives the 1990 distribution from *Chess Life and Review* ("1990 Annual Rating List," 1991) The mean is 1577 (*SD* = 362) This distribution is much less skewed, though variability seems to have increased strikingly (As we discuss later, this increase in variability can lead to very large differences in the expected value of the best player) Thus, to a first approximation, the distribution of chess ratings (based on tournament chess performance) is reasonably well described by a normal distribution As Elo (1986) pointed out, a Maxwell-Boltzmann distribution does a slightly better job of fitting the 1977 USCF data than does a normal distribution

Tests of the Elo (1986) rating system show it to be remarkably valid for predicting tournament and world-championship results Batchelder and Bershad (1979) and Henery (1992) have provided a formal development of the Thurstone-Mosteller scaling model to support Elo's empirically driven rating scale

SEX DIFFERENCES IN CHESS PERFORMANCE DUE TO PARTICIPATION RATES?

FIDE has held separate men's and women's tournaments and publishes separate rating lists of the best men and best women players There seems to be little a priori justification for segregation by sex for an intellectual pursuit such as chess Perhaps the rationale is that historically the best women have performed far worse than the best men

There is considerable speculation about the reason for the "poor" performance of women in chess, ranging from biological notions of spatial ability being inferior in women (summarized in Holding, 1985) to specific hypotheses that spatial ability is depressed by menstrual cycle changes (e g , Chabris & Hamilton, 1992) Holding (1985) appeared to side with a socialization view that women may be raised to be less competition oriented But before concluding that there are underlying differences in native ability to plan out chess moves, we ought to examine participation rate as an explanation

Women members represent 4 5% of the USCF population (Lawrence, 1993), for a male female ratio of 21 1 Berry (1993) indicated that women form about 2% of the membership of the Chess Federation of Canada (a 49 1 ratio) The 1990 FIDE rating list of the 60 nations with both men and women players demonstrates a much smaller ratio, with a mean ratio of 10 7 and *SD* of 8 8 The range runs from 42 1 to 0 9 1 (Italy with 127 men to 3 women, Mongolia with 11 women to 10 men) But there are many small countries with no FIDE-rated female players The ratio tends to rise as the number of male players increases, so that when there are 30 or more men, the mean ratio is 14 1

If the male female ratio is approximately 20 1, then by the log-linear *max* relationship, the best male player in the world will be $\log_{10}$ (20) × 0 66 *SD* units better than the best female player If the *SD* for rating points is 250, the best male player will be rated 215 points higher than the best female player 1 3 × 0 66 × 250 rating points = 215 rating points

If the male female ratio is narrower, say, 11 1, then the best male player will be superior by 172 rating points (1 04 × 0 66 × 250)

On the January 1994 FIDE rating lists, the best man was Gary Kasparov, with 2805 Elo points (his 1993 rating, because he was not officially ranked in 1994 after splitting with FIDE), and the best woman was Judit Polgar, with 2630 Elo points The difference of 175 rating points is within the range predicted by ratios of 11 1 to 20 1

Thus, we should not reject the hypothesis that women and men come from the same underlying distribution of chess skill Admittedly, failing to reject the null does not constitute a strong result, and the difference could also be explained by some other combination of assumptions Our claim is that MILL7 provides a good baseline explanation against which others can be evaluated That is, it is a better null hypothesis than the typical nil one (Cohen, 1994) of a mean difference of zero

It is logically possible to explain the gap in top performance (or top performance based on fractions of the top *m* performers) based on equal means but different standard deviations among men and women players, a point made by Becker and Hedges (1988) and Humphreys (1988) in their discussion of the sex difference favoring males in elite Scholastic Aptitude Test (SAT) mathematics performance (reported by Benbow, 1988)

Another way to provide converging evidence for the log-linear relationship between population size and extreme scores in chess is to plot the best rated player for each country, separately for men and women, against the country's participation rates Figure 3 provides the regressions for men and women of the maximum Elo rating by country (from the 1990 FIDE lists) as a function of the estimated log number of rated chess players in that country (from 1982 figures provided by Elo, 1986) We

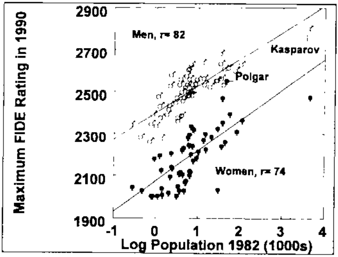

Fig 3 Rating of the top player by log size of the country's chess-playing population for 55 countries Results for men and women are shown separately FIDE = Fédération Internationale des Echecs

use a time-lag comparison because Simon and Chase (1973) argued that it takes about 10 years to develop chess ability to grand-master levels. The plot shows reasonably good linear relationships between maximum rating and log participation rate, and essentially parallel lines for men (top) and women (bottom).

We take the near equivalence of the slopes for men (128, *SE* = 12) and for women (144, *SE* = 18) as evidence that the log-linear relationship is essentially the same in men and women. There is a significant difference in the intercepts of 324 rating points (2400 for men and 2076 for women, with *SE* = 12.6 and 18.7, respectively). This difference may reflect the difference in participation rates between men and women (the data are not available), or a difference in the mean levels of performance between these top men and women, or perhaps different variability in these populations, or some combination of these factors.

We can make a rough test of the difference between the top ratings within countries. The difference between the top man and woman expressed in *SD* units, divided by the log ratio of the participation rates, should be a constant of 0.66. We divided the rating difference by 250 (the USCF *SD*) to create the variable *DIFFSD*, and took the log (base 10) of the ratio of men to women among FIDE-rated players (*LOGRATIO*). We created the constant CON66 = *DIFFSD/LOGRATIO*. The mean for CON66 (*N* = 60 countries) was 1.4, with an *SE* of 0.37, yielding a 95% confidence interval of 2.14 to 0.66. Although the predicted value, 0.66, is just within the confidence band, we take this as rather weak evidence. The problem is that male:female ratios of small countries with very few participants (players with FIDE ratings) are weighted as heavily as the more stable ratios of large countries. The estimate that is probably the most reliable is the one for the United States, where the *SD* of ratings is known to be about 250 and the ratio of men to women is known to be about 20:1. In the FIDE-rated sample, the male:female ratio for the United States is 18.2:1. The value of CON66 for the United States is 0.68.

PREDICTING INTERNATIONAL DIFFERENCES IN *MAX* FOR CHESS BY POPULATION SIZE

As Elo (1986) pointed out, about a third of all the world's grand masters are from the (former) Soviet Union. A grand master (rating of ~2500+ Elo points) is approximately 4 standard deviations above the mean ($Z = 3.8$). Elo attributed this dominance of the Soviet Union in chess to training (the so-called Soviet School of Chess) and "the statistical fact that every bit of talent is developed, an example of a higher pyramid peak when the pyramid base is broadened" (p. 111). He also noted that "the fraction [of grand masters] would be even larger if more Soviet masters competed outside the Soviet Union" (p. 110).

We can evaluate Elo's seemingly paradoxical observations about the Soviet Union indirectly by determining if the strongest grand master is at a level above or below what would be expected under MILL7 via Equation 1. Our best estimates for a world mean and standard deviation for chess come from the known properties of the USCF rating distribution, with a mean for established players of roughly 1550 and a standard deviation of roughly 250.

Elo (1986) gave the number of registered players in the Soviet Union in 1982 as 4 million. From Equation 1, the best player should have approximately $1550 + 0.66 \times \log_{10} (4{,}000{,}000) \times 250 = 2640$ rating points. Kasparov's rating was 2805, considerably above this estimate. We can conclude either that the Soviet School of Chess works or that the estimates we used for μ and σ are in error.

If we estimate σ at 300, a value closer to the mean of the 1977 and 1990 USCF values, the estimated peak score is 2857, showing that *max* is quite sensitive to σ. Another way to converge on the problem is to use the regression line for *max* as a function of the size of the chess-playing population; this regression line uses all available data for *max* and country populations. For the data for men from Figure 3, the regression equation is

$$max = 2400 + 128 \times \text{Log (Population in 1000s)}$$

For the former Soviet Union, this equation predicts a *max* of $2400 + 128 \times 3.6 = 2861$, a value pretty close to that achieved with $\sigma = 300$. Kasparov is possibly a bit underrated by these estimates, though close enough that he is not out of line. The weight of evidence is consistent with a view that the Soviet training system does not exert an effect beyond what could be predicted from the participation rate in chess in the general population. (Paradoxically, Bobby Fischer's ascent to a similar rating level in 1972 could lead to the interpretation that the "American System" is superior to the Russian one, given the smaller U.S. chess population.)

CAVEATS

MILL7 is a useful baseline measure for assessing differences in maximal (or minimal) performance. Its application could prevent a misleading inference when group differences for elite performers could really be merely a function of differences in participation rate, as appears to be the case in women's chess performance. There are, however, many correlated assumptions in applying this measuring stick. Foremost is the assumption that everyone comes from the same normally distributed population. Also, one assumes that there is equally efficient transport of top performers through the system. A mixture of differences in μ and in σ within subgroups could also invalidate the inference process. Another potentially worrisome assumption is that the estimate of the population standard deviation is not already contaminated by differential training effects in the comparison groups.

SOME SOCIAL IMPLICATIONS OF THE MILL7 FUNCTION

There are some interesting corollaries to the MILL7 function. Assume that a population consists of the majority group, Major, and a minority group, Minor. Assume that Minor applies for employment in proportion to their numbers in the population. Assume further that Minor and Major come from the same distribution of ability (same μ and σ). A quota or affirmative action system that chooses *max* of Minor over *max* of Major can expect to find the former always performing worse than the latter by

$$\log_{10}((p(\text{Major})/p(\text{Minor})) \times 0.66\ SD$$

For instance, if there are 10 times as many Majors as Minors in the population, *max* of Minor can be expected to be about 0.7 *SD* units worse than *max* of Major, under the assumption of no differences in the underlying ability for the two groups. What MILL7 provides is an estimate of how inferior that minority candidate can be expected to be. It serves as the baseline for the expected result of quota hiring. When there are objective measures of quality that can be phrased in *SD* units, MILL7 and the web of accompanying assumptions can be tested.

BLACK-WHITE AND MALE-FEMALE DIFFERENCES IN MAXIMAL PERFORMANCE

It may also be useful to frame other intergroup differences in a MILL7 context. If whites and blacks attend equally well to educational instruction at roughly their proportions in the U.S. population (about 8:1), then a 0.6 *SD* difference is to be expected for the top academic performers from the two subgroups, assuming that there are no differences in mean or standard deviation favoring the majority group. Recall that Figure 1 showed that the slope of the log-linear function increases for comparisons of people further down the rank listing. Thus, we expect that comparisons of small fractions of top-ranked whites against blacks (e.g., Herrnstein & Murray, 1994, comparing top scorers on the Medical College Admissions Test, Law School Admissions Test, and Graduate Record Exam) would yield even larger gaps in achievement, even if the population means and standard deviations were equal.

Similarly, if among mathematically talented preadolescents the ratio of elite males to elite females is 5 to 1, we would expect differences in the very top scorers of at least 0.5 *SD*. Hyde, Fennema, and Lamon (1990) favored participation selectivity as the explanation of the larger effect size for sex differences in SAT mathematics performance (Benbow, 1988) than in general mathematics performance.

Because most tests of differences are not tests on the whole population but tests on a select subgroup, it is likely that some of the difference in performance will be due to the differences in the top *m* percentiles of the two distributions, rather than solely to differences in the population means. It would be ideal to have slope estimates for the expected values of the means for top percentiles rather than for individual ranks.

NATIONAL TRAINING STRATEGIES

We can tentatively offer some strategies for winning Olympic medals or world championships. The general strategies fall into three categories and their hybrids. One strategy is to allocate resources to training a few elite performers, hoping to push them up to world-class levels. This is equivalent to maximizing σ, holding *N* and μ constant: the *large σ strategy*. Another option is to try training everyone in the population a little bit, to push up μ, but leave *N* and σ unchanged: the *large μ strategy*. Or one can encourage mass participation, providing participation incentives but no training to the masses, relying on MILL7 to float *max* (hidden somewhere in that great untapped sample) to the top. This is equivalent to maximizing *N* and holding σ and μ constant: the *large* N *strategy*. A final possibility is to engage in some hybrid of these strategies (Gerchak & Kilgour, 1994). Which route is most effective depends in part on whether external training is the most important factor in skill development.

For example, in chess (Simon & Chase, 1973) or music (Ericsson, Krampe, & Tesch-Romer, 1993), external training has only modest effects, and experts mostly train themselves through thousands of hours of deliberate practice (Ericsson & Charness, 1994). Although coaching and formal training are important in these domains, as Bloom (1985) emphasized, the relative amount of time spent with external coaches is very small (hundreds of hours?) compared with individual practice time (measured in thousands of hours). Perhaps a good rule of thumb is a 10:1 ratio of individual to instructed practice as a break point for calling an activity self-training rather than external training. The nation with the most participants in such an endeavor will usually have the world champion. Large *N* is probably the preferred strategy. Recruitment is probably cheaper than individual training.

In contrast, when training is crucial (e.g., education, Nobel Prize in a technically demanding area), the large σ and large μ strategies seem better. In such cases, it is probably more effective to increase σ (add a constant to a fraction of the population) or to try to raise the mean (add a constant to everyone's score) than to increase *N*.

The trade-off between training a subset (the elite) or training everyone is unclear because we cannot easily estimate how much the mean and standard deviation change under these two policies. It is clear that increasing the variability (standard deviation) pays greater dividends when you deal with very extreme performances. Such an emphasis on elitism is, however, often socially unacceptable in domains such as public education, where the goal is usually to move each member of society to acceptable performance levels rather than to turn out a few high performers at the expense of the majority.

CONCLUSIONS

Before it is possible to conclude much about national, gender, and racial differences in performance on tasks such as chess playing, one needs to take participation rates into account. As we have shown, very large individual differences can arise from each 10-fold increase in group size. The MILL7 model provides a convenient method for estimating changes in maximal performance. Considerably more work is needed to check whether the performance differences of, say, the 20th best male and female, or Russian and non-Russian, chess players, and other differences in rank, will also be reasonably well explained by differences in the number of participants alone.

Needless to say, even if the difference in the number of keen participants in an activity can account for the difference in achievements, the question remains why the numbers or proportions of keen participants vary so sharply across some groups. Why do so many men in some countries, and not others, play chess? Why do so many American youths, especially blacks, pursue basketball seriously? If the number of participants itself happens to reflect true or perceived differences in innate abilities, this fact is clearly not captured by our model Nonetheless, we believe that MILL7 provides a computationally simple baseline model worth testing before moving on to alternative explanations

Acknowledgments—Support for this research has been provided to the first author by the Natural Sciences and Engineering Research Council of Canada, NSERC A0790, the Canadian Aging Research Network, one of 15 Networks of Centres of Excellence supported by the government of Canada, the German government's Deutscher Akademischer Austauschdienst leave fellowship, and the Max Planck Institute for Human Development and Education, Berlin Funding for this research has been provided by the Natural Sciences and Engineering Research Council of Canada NSERC OGP0005698, to the second author We are grateful to Cathy Kelley for running the Monte Carlo simulations and helping with some of the data analysis We also thank John Nesselroade, Alan Hartley, and anonymous reviewers for helpful comments

This article is dedicated to the memory of Arpad E Elo (1903–1992), whose chess performance rating scale (Elo, 1986) has been an invaluable tool for research on chess skill

REFERENCES

Arnold B C Balakrishnan N & Nagaraja H N (1992) *A first course in order statistics* New York Wiley

Batchelder, W H & Bershad N J (1979) The statistical analysis of a Thurstonian model for rating chess players *Journal of Mathematical Psychology 19* 39–60

Becker B J & Hedges, L V (1988) The effects of selection variability in studies of gender differences *Behavioral and Brain Sciences 11* 183–184

Benbow C P (1988) Sex differences in mathematical reasoning ability in intellectually talented preadolescents Their nature, effects, and possible causes *Behavioral and Brain Sciences 11* 160–232

Berry, J (1993 July 31) Women players Results puzzling *The Globe and Mail*, p C2

Bloom, B S (Ed) (1985) *Developing talent in young people* New York Ballantine Books

Chabris C F & Hamilton, S E (1992) Hemispheric specialization for skilled perceptual organization by chessmasters *Neuropsychologia 30*, 47–57

Cohen J (1994) The earth is round ($p <$ 05) *American Psychologist*, *49* 997–1003

Elo A E (1986) *The rating of chessplayers past and present* (2nd ed) New York Arco

Ericsson K A & Charness N (1994) Expert performance Its structure and acquisition *American Psychologist 49* 725–747

Ericsson K A Krampe R Th & Tesch-Römer, C (1993) The role of deliberate practice in the acquisition of expert performance *Psychological Review*, *100* 363–406

Gerchak Y & Kilgour D M (1994) *Optimal parallelism in R&D and sports funding How many baskets for the eggs?* Manuscript submitted for publication

Harter, H L (1961) Expected values of normal order statistics *Biometrika 48* 151–165

Henery R J (1992) An extension of the Thurstone-Mosteller model for chess *The Statistician 41* 559–567

Herrnstein, R J & Murray C (1994) *The bell curve Intelligence and class structure in American life* New York Free Press

Holding D H (1985) *The psychology of chess skill* Hillsdale, NJ Erlbaum

Humphreys L G (1988) Sex differences in variability may be more important than sex differences in means *Behavioral and Brain Sciences 11* 195–196

Hyde J S Fennema E & Lamon S J (1990) Gender differences in mathematics performance *Psychological Bulletin 107* 139–155

Joshi P C & Balakrishnan N (1981) An identity for the moments of normal order statistics with applications *Scandinavian Actuarial Journal 47*, 203–213

Lawrence A (1993) Analytical mistake *Chess Life 48*(3), 6

Logan G D (1988) Toward an instance theory of automatization *Psychological Review 95* 492–527

1990 Annual Rating List (1991) *Chess Life and Review 46*(1) 73

Simon H A & Chase W G (1973) Skill in chess *American Scientist 61*, 394–403

Tippet L H C (1925) On the extreme individuals and the range of samples taken from a normal population *Biometrika 17* 364–387

(RECEIVED 9/27/94, REVISION ACCEPTED 1/28/95)