Anyone for chess?
Analysing chess ratings above high thresholds

Nils Lid Hjort

Department of Mathematics, University of Oslo

January 7, 2026

Abstract. Suppose some cleverness score parameter is sufficiently interesting to be defined and then measured, perhaps for different strata of specialists or for the broader population. Such phenomena could have Gaussian distributions, when it comes to all players in a stratum, but when interest focuses on the very tails, for the top few percent, those above certain high thresholds, different models are called for, along with the need to analyse such based on the listed top scores only. In this note I develop such models and tools, and apply them to the top-100 and above 2100 points lists for regular chess ratings, for the currently active 14671 men and 753 women, as given by the FIDE, January 2026. It is argued that even when two or more distributions have close to identical expected values, or medians, even smaller differences in variance may explain gaps for the few very best ones.

Key words: chess ratings, gender gap, from tails to the main bulk, Magnus Carlsen, models for scores above thresholds, participation volume, variance ratio

1 Introduction: the sizes of maxima

Figure 1: Queen's Gambit.

The beautiful game of chess is for nearly everyone: just grab the pieces and your nearby board, on your laptop or on your phone, or perhaps your analogue bishops and rooks and knights and pawns, and play on. It takes beautiful minds to play it really wondrously, however. "Not In 10,000 Lifetimes Will We See Chess Like This Again", proclaims *Epic Chess*, among the many chess experts on the net, with hundreds of thousands of followers, analysing in detail Magnus Carlsen's complicated march towards winning the World Rapid and World Blitz, in Doha, Christmas week 2025. So statisticians should be interested too.

In this note I study the chess ratings, as per the Elö scale used by the International Chess Federation FIDE in clubs and tournaments all over the world, not on the population scale but for the very clever ones, those with ratings above a high threshold, like $r_0 = 2100$. Before I come to chess, however, with models and analysis methods developed in Section 2, it is useful to build some intuition around extreme scores, under natural and simple sets of assumptions. The aim is partly to identify what the driving factors are, for extreme scores inside different strata.

Consider therefore M_n, the maximum of $X_1, \ldots, X_n$, an i.i.d. sample from the $\mathrm{N}(\xi, \sigma^2)$. We may think in terms of a stratum of people and their heights, with M_n the height of the tallest. Then, with $c_n \doteq d_n$ indicating $c_n/d_n \to 1$,

$$\mathrm{E}\, M_n \doteq \xi + \sigma(2 \log n)^{1/2},$$

which captures the essential ways in which the tallest can be very tall (or perhaps not so very tall, after all): (i) ξ should be big; (ii) higher participation, i.e. a bigger n, matters, but only on the very slowly growing $(\log n)^{1/2}$ scale; and (iii) it is also helpful with a not small σ. There is a similar approximation also in the equicorrelation case, where individuals in a stratum or population have 'something in common', modelled as $\mathrm{corr}(X_i, X_j) = \rho$ for all $i \neq j$. In that case,

$$\mathrm{E}\, M_n \doteq \xi + (1-\rho)^{1/2} \sigma (2 \log n)^{1/2}, \tag{1.1}$$

so with more similarity than under pure independence, the maxima follow the same pattern, modulo a reduction factor; more independence between individuals imply bigger extremes.

For illustration, consider countries A and B where something of sufficient interest is recorded for their people have distributions $\mathrm{N}(\xi, \sigma_A^2)$ and $\mathrm{N}(\xi, \sigma_B^2)$. The 'winners' from A and B can be expected to have scores

$$\xi + \sigma_A (2 \log n_A)^{1/2} \quad \text{and} \quad \xi + \sigma_B (2 \log n_B)^{1/2},$$

in terms of population sizes n_A and n_B. The excess ratio, B compared to A, is essentially

$$r = \frac{\sigma_B (\log n_B)^{1/2}}{\sigma_A (\log n_A)^{1/2}}.$$

For the participation ratio, note that if A is e.g. 3 times bigger than B, the sample size ratio is

$$\frac{(\log 3 + \log n_B)^{1/2}}{(\log n_B)^{1/2}} = \Big(1 + \frac{\log 3}{\log n_B}\Big)^{1/2},$$

indicating that the difference matters, when it comes to the highest-scoring levels, but not by much, when it comes to very high sizes. The tallest Dutch cannot be expected to be much taller than the tallest Norwegians, assuming here these countries have the same grand average and the same grand variances, even if the Duch population s three times that of the Norwegians. Neither is it entirely impossible that a guy from Lommedalen is taller than all Russians.

In many cases, therefore, the standard deviation ratio matters rather more than the sample sizes. To get even with B, A needs

$$\frac{\log n_B}{\log n_A} = \frac{\sigma_A^2}{\sigma_B^2}, \quad \text{or } n_A = n_B \exp(\sigma_B^2/\sigma_A^2),$$

featuring the crucial variance ratio. For illustration, suppose A and B have $\mathrm{N}(100, 15.0^2)$ and $\mathrm{N}(100, 16.0^2)$; then B will indeed tend to have the tallest scores. Country A needs to be 15 percent bigger than B in order to make up for the small 6 percent standard deviation increase.

2 A parametric model for ratings over threshold

My two-parameter model for scores above threshold has density of the form

$$f(x, a, \theta) = \Gamma(a+1)^{-1} \exp\Big\{-\Big(\frac{x - r_0}{\theta}\Big)^{1/a}\Big\}\frac{1}{\theta}, \quad \text{for } x \geq r_0, \tag{2.1}$$

monotonically decreasing on its domain. We shall see in Section 3 that the fit to chess scores data is very good. Here a is driving the fine aspects of the upper tail, whereas θ is a scale parameter reflecting the variability of the distribution.

2.1 Properties. A little exercise verifies that the integration constant in (2.1) is the right one, and that the model also can also be represented as

$$X = r_0 + \theta\, V_a^a \quad \text{where } V_a \sim \mathrm{Gamma}(a, 1). \tag{2.2}$$

Formulae for moments and quantiles may be deduced from this. In particular,

$$\mathrm{E}\, X = r_0 + \theta\frac{\Gamma(2a)}{\Gamma(a)}, \quad \mathrm{Var}\, X = \theta^2\Big[\frac{\Gamma(3a)}{\Gamma(a)} - \Big\{\frac{\Gamma(2a)}{\Gamma(a)}\Big\}^2\Big]. \tag{2.3}$$

Furthermore, solving $F(x, a, \theta) = p$ for x leads to the quantile function

$$F^{-1}(p, a, \theta) = r_0 + \theta\, G_a^{-1}(p)^a. \tag{2.4}$$

Also useful is an explicit formula for the cumulative distribution function (c.d.f.), via (2.2):

$$F(x, a, \theta) = P(r_0 + \theta\, V_a^a \leq x) = G_a(((x - r_0)/\theta)^{1/a}) \quad \text{for } x \geq r_0, \tag{2.5}$$

in terms of the c.d.f. G_a for the $\mathrm{Gamma}(a, 1)$.

The maxima of samples of players from some given stratum, like all active players in a country, follow the pattern of (1.1), though with an adjusted logarithmic speed, depending on the a. Specifically, with M_n the maximum of an i.i.d. sample from (2.1), we have

$$\mathrm{E}\, M_n \doteq r_0 + \theta(\log n)^a \tag{2.6}$$

for growing participation volume n. The Gaussian case corresponds to $a = \frac{1}{2}$ and the exponential to $a = 1$; the chess scores have tails between these two classical cases. Formula (2.6) follows from extreme value theory; we may point to Gumbel (1919, 1922, 1958), Embrechts et al. (1997). There is a Gumbel distribution limit for the maximum here, with $(M_n - a_n)/b_n \to_d G$, the Gumbel distribution with c.d.f. $\exp(-\exp(-g))$, for suitable slowly increasing a_n and b_n, with consequences also for all quantiles; see the details in Remark B of Section 5.

2.2 Estimation and inference. Consider a stratum of interest, with ratings $x_1, \ldots, x_n$ for its n individuals. We need to estimate the parameters (a, θ) in the (2.1) model.

Suppose first that the full set $x_1, \ldots, x_n$ is available. The log-likelihood is then

$$\ell(a, \theta) = \sum_{i=1}^{n} \Big\{ -\log \Gamma(a+1) - \Big(\frac{x_i - r_0}{\theta}\Big)^{1/a} - \log \theta \Big\}, \tag{2.7}$$

which can be maximised numerically. When needed, for numerical precision or for bigger datasets, the maximisation can be reduced to a function of a alone, namely $\ell(a, \widehat{\theta}(a))$, as the log-likelihood for given a is maximised for

$$\widehat{\theta}(a) = n^{-1} \sum_{i=1}^{n} (x_i - r_0)^{1/a}.$$

Intriguingly and importantly, parameters may be estimated also in cases where only the say the top-100 ratings are listed; in various databases for chess and performance sports only such top-k lists are available, for suitable cutoff k. Sorting these top k values as $x_1 > \cdots > x_k$, then, one merely knows that the other $n - k$ ratings are inside the $[r_0, x_k]$ box. The statistical information, clearly less complete compared to knowing all the data, can be represented by the likelihood

$$L_k(a, \theta) = F(x_k, a, \theta)^{n-k} f(x_k, \theta) \cdots f(x_1, \theta),$$

i.e. with log-likelihood

$$\ell_k(a, \theta) = (n-k) \log F(x_k, a, \theta)^{n-k} + \sum_{i=1}^{k} \log f(x_k, a, \theta), \tag{2.8}$$

which again can be maximised to fit the available top-k dataset. This takes knowing the participation volume, i.e. the n, though.

3 Chess ratings, for men and women, per January 2026

From the FIDE website I have got hold of the 15,424 players, active as of January 2026, with regular ratings equal to or above the threshold $r_0 = 2100$, with $n_m = 14671$ male and $n_w = 753$ female players. Top-rated Magnus Carlsen has 2840, whereas Hou Yifan with 2613 leads the women table; we note, for further analysis below, that the top-100 rated players are all men.

We can now fit the parameters of the (2.1) density by maximising the log-likelihood (2.7), for the two categories. Approximate standard deviations may also be computed, for the basic model parameters and for smooth functions thereof, via the delta method; see e.g. Hjort & Stoltenberg (2026, Chs. 2, 5). In some detail, for the maximum likelihood (ML) estimators,

$$\begin{pmatrix} \widehat{a} \\ \widehat{\theta} \end{pmatrix} \approx_d \mathrm{N}_2\Big(\begin{pmatrix} a \\ \theta \end{pmatrix}, \widehat{J}^{-1}\Big),$$

with $\widehat{J}$ the observed Fisher information matrix, i.e. minus the Hessian at the ML position. Also, for any focus parameter $\mu = \mu(a, \theta)$, like the mean, the standard deviation, or a quantile, we have

$$\widehat{\mu} = \mu(\widehat{a}, \widehat{\theta}) \approx_d \mathrm{N}(\mu, \widehat{\tau}^2), \quad \text{with } \widehat{\tau}^2 = \widehat{c}^{\,\mathrm{t}} \widehat{J}^{-1} \widehat{c},$$

where $\widehat{c}$ is the gradient of the μ evaluated at the ML. Results are as follows, for the two strata, with estimates and standard errors (estimated standard deviations):

```
Table I:
         men                 women
a          0.689 0.013        0.612  0.052
theta    209.28  3.49       194.86  12.74
mu      2241.52  0.99      2221.53   3.60
sigma    119.67  1.07        98.31   1.58
median  2210.79  1.05      2198.23   3.98
```

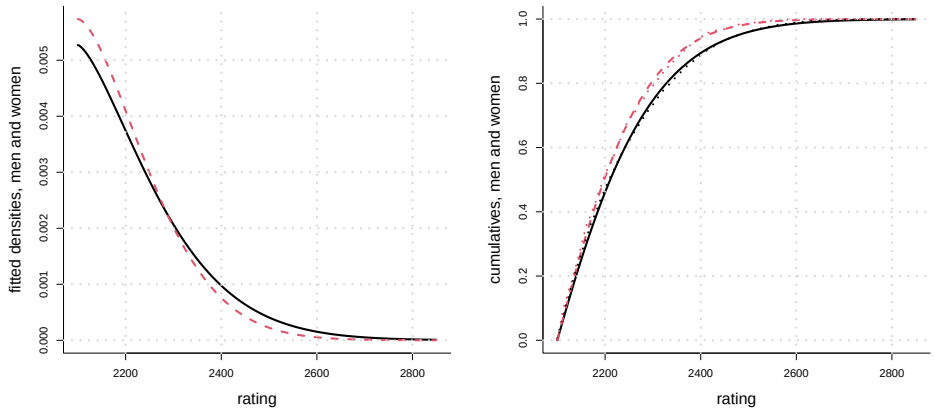

Figure 2: Left panel: fitted densities $\widehat{f}_m$ and $\widehat{f}_w$, for men (full, black) and women (red, slanted) having rating above 2100; they are not far apart but the right tail is dominated by the men. Right panel: fitted c.d.f.s $\widehat{F}_m$ and $\widehat{F}_w$, along also with the empirical c.d.f.s (very slightly non-continuous, as dotted curves). The parametric and nonparametric are very close, i.e. the model used is fully adequate.

The estimated densities $\widehat{f}_m$ and $\widehat{f}_w$ for men and women are displayed in Figure 2, left panel; these are seen to be close for the main bulk of players, but with a significantly fatter right tail for the men, That aspect, which shows up also for c.d.f.s and quantile functions, is not coupled with the higher participation number for men, per se, but by the upper performances of the two groups. A mild caveat, though, is that the estimated curves and quantities for the women group are less precise than those for the men.

In Figure 2, right panel, both the parametric and empirical c.d.f.s are portrayed, i.e.

$$\begin{aligned} \widehat{F}_m(x) &= F(x, \widehat{a}_m, \widehat{\theta}_m), & \widehat{F}_w(x) &= F(x, \widehat{a}_w, \widehat{\theta}_w), \\ F_{\mathrm{emp},m}(x) &= n_m^{-1} \sum_{i=1}^{n_m} I(x_{m,i} \le x), & F_{\mathrm{emp},w}(x) &= n_w^{-1} \sum_{i=1}^{n_w} I(x_{w,i} \le x). \end{aligned}$$

We learn from this that the model (2.1) is even surprisingly adequate; formal monitoring of $\widehat{F}_w - F_{\mathrm{emp},w}$ shows no significant deviation from zero, etc.

In Figure 3, left panel, we also plot the difference $S_{\mathrm{emp},m} - S_{\mathrm{emp},w}$ (black, wiggly), with $S_{\mathrm{emp},m}(x)$ and $S_{\mathrm{emp},w}(x)$ defined as the the remaining factions above x. So at all levels x the male proportion above it is a few percent more than for the femails. The red slanted curve is the corresponding parametrically estimated difference, again indicating a very good fit. The lower and upper blue wiggly curves form a 90 percent confidence band for the

$S_m - S_w$ difference, indicating also that the differences noted are statistically significant. Finally, that figure's right panel displays the estimated quantile functions, again with the parametric fits very close to the empirical ones. Visibly and significantly, there is a gender gap at the very top levels (for the populations studied, those above 2100 regular rating, for those active as of January 2026).

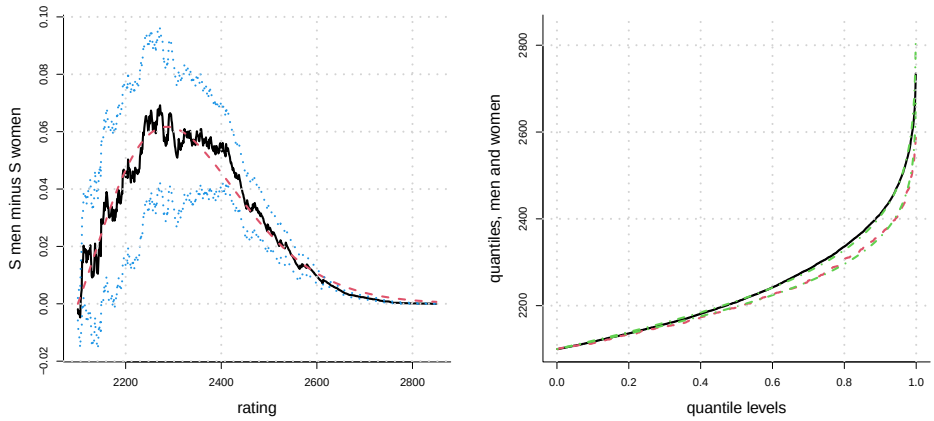

Figure 3: Left panel: the difference $S_m(x) - S_w(x)$, empirical (black, wiggly) and parametrically fitted (red, slanted, smooth), along with a 90 percent confidence band. Right panel: empirical and theoretical quantile functions Q_m (black, full) and Q_w (red, slanted). The model fit is excellent.

4 Assessing the gender gap

We have seen above, see e.g. Table I, that the $a_m = a_w$ hypothesis is not in conflict with the data. For the two standard deviation estimates 119.67 and 98.31, however, with standard errors 1.07 and 1.48, we have

$$t = \frac{\widehat{\theta}_m - \widehat{\theta}_w}{(\mathrm{se}_m^2 + \mathrm{se}_w^2)^{1/2}} = \frac{21.26}{1.91},$$

indicating that the male variance is much stronger than the female one. This also agrees with our findings and figures above. Whereas Figure 2, left panel, shows the estimated direct densities $\widehat{f}_m, \widehat{f}_w$, Figure 4, left panel, shows their logarithms, i.e. $\log \widehat{f}_m$ and $\log \widehat{f}_w$, showing more clearly that the male tail is bigger and fatter than the female tale.

We hence already know well that there are significant differences between the top men and top women players, for regular chess ratings, as of January 2026. It is instructive to build other formal tests as well. A class of such tests take the following form, using bootstrap samples. Pooling the men and women together, with $n = n_m + n_w = 15424$ chess scores, sample randomly $n_m = 14671$ of these, labelled x_m^*, the remaining $n_w = 753$ labelled x_w^*. From such a random bootstrap split of the full dataset, we may compute discrepancy scores, like

$$A = (F_{\mathrm{emp},m}^*)^{-1}(0.90) - (F_{\mathrm{emp},w}^*)^{-1}(0.90) \quad \text{and} \quad B = \mathrm{sd}(x_m^*) - \mathrm{sd}(x_w^*),$$

the differences of 0.90 quantiles and of standard deviations. Doing this say 1,000 times yields bootstrap null distributions of A and B, under the no-difference hypothesis. This is portrayed in Figure 4, right panel. Then we compute the actually observed $(A_{\mathrm{obs}}, B_{\mathrm{obs}}) =$

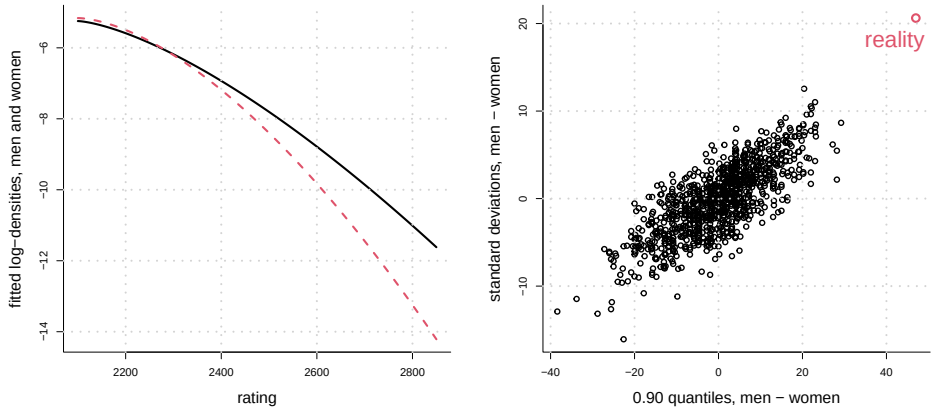

Figure 4: Left panel: the estimated log-densities $\log \widehat{f}_m$ and $\log \widehat{f}_w$, indicating that the man tail is fatter and longer than the woman tail; more top players are, indeed, men. Right panel: bootstrap points (A^*, B^*) for 1,000 bootstrap null distribution datasets, along with the factual $(A_{\rm obs}, B_{\rm obs})$, for A the 0.90 quantile difference and B the standard deviation difference.

$(47.0, 20.6)$, plotted in the right up corner of that figure. We see again, in a clear fashion, that the null hypothesis of gender equality is firmly rejected; the differences $A_{\rm obs}$ and $B_{\rm obs}$ are far outside the null hypothesis range.

5 Concluding remarks

In this note I have introduced a two-parameter model for scores above thresholds, and demonstrated that it very adequately describes the statistics of top players in chess. The model, with relevantly estimated parameters, may be used to monitor progress over time, to predict overall level performances in tournaments, and can surely be applied to other top scores phenomena, in biology or otherwise.

To round off this note a few comments are as follows.

Remark A. Same tail index for men and women. Slightly sharper estimates might be given, than those obtained in Table I of Section 3, via estimation in the three-parameter model which takes $a_m = a_w$; model selection analysis, with the AIC and the BIC, cf. Claeskens & Hjort (2008, Chs. 2,3), indicate that this three-parameter reduction has a slightly better explanatory power than needing two plus two parameters. It is also of interest to see in this fashion that the tail parameters are quite close for the two groups (the null hypothesis $a_m = a_w$ is not refuted), indicating that the primary factor underlying the current gender gap in top chess is not participation, or different skills for the main bulk of top players, but variability.

Remark B. Sample extremes. Suppose $u_1, \ldots, u_n$ are an i.i.d. sample for the density $f(u) = \Gamma(a+1)^{-1} \exp(-u^{1/a})$ on the halfline, with tail behaviour dictated by the a parameter. For large u, the survivor function is

$$\bar{F}(u) = 1 - F(u) \doteq c u^{1-1/a} \exp(-u^{1/a}),$$

with the dominant term being the stretched exponential. The tail is Gaussian for $a = \frac{1}{2}$ and exponential for $a = 1$. The sample winner $M_n = \max_{i \le n} u_i$ will have $\mathrm{E}\, M_n \doteq (\log n)^a$.

More precisely, there is a Gumbel limit,

$$(M_n - b_n)/a_n \to_d G, \quad \text{with } a_n = a(\log n)^{a-1},\ b_n = (\log n)^a,$$

which leads to

$$\mathrm{E}\, M_n = (\log n)^a + \gamma_e a(\log n)^{a-1} + o((\log n)^{a-1}),$$

with $\gamma_e = 0.5772...$ the Euler constant.

For our chess ratings model, then, in view of the representation (2.2), the expected top rating, for a stratum with n players above 2100, is

$$\begin{aligned} \mathrm{E}\, M_{n,m} &\doteq 2210 + 210.22\,(\log n)^{0.68}, \\ \mathrm{E}\, M_{n,w} &\doteq 2200 + 179.04\,(\log n)^{0.61}, \end{aligned}$$

for men and women. These estimates, or predictions for future years or other strata, can be sharpened somewhat in view of Remark A.

Remark C. The Rashomon Syndrome. Statistical modelling and inference can be very good at explaining the statistical essence of what goes on for some perhaps complicated phenomenon studied, in terms of pointing to mechanisms, predicting what happens next, spotting and assessing differences between groups, etc. etc. A sound caveat is that of the so-called Rashomon Syndrome, however; rather different statistical 'explanations', from perhaps different models and arguments, might turn out to be explaining the data equally well. So even though the (2.1) model succeeds very well in fitting all such higher ratings chess data, it should not tbe seen as the canonical 'explanation' for what goes on in the minds of the 15,000 most clever chess players of our time.

Remark D. Learning the full iceberg from only seeing the top. As explained with (2.8), the full model can be learned from recording say only the top-100, as long as we also know the participation level, i.e. the size n above the threshold in question. This should be useful when it comes to analysing chess ratings for various strata or groups for which the full datasets are not fully known.

Remark E. Other gender gap literature. There are many articles and stories, published in journals or as blog posts or expert views in social media, pertaining to the gender gap in chess; the extent to which it is present, what likely causes might be, and how to amend, in the direction of less imbalance. A simple search on Google Scholar finds more than a hundred publications. Good articles published on the web include *What gender gap in chess?*, by Wei Ji Ma, who argues that participation is the dominating factor. The present note does not disagree with this, but points to the crucial presence of *statistical variation* within strata.

Remark F. Application to other domains. Clearly, 'high scores above threshold' extreme value data abound in other application areas, from finance and meteorology and sports, see e.g. Hjort (2026), to biological traits, like heights and academic skills. Note e.g. via Remark D that if we know the tallest 50 US basketball players, in a season with 500 players, we may actually infer the full distribution.

Acknowledgments

I greatly enjoyed following the Norwegian Broadcasting Corporation's home channel televising the World Rapid and World Blitz from Doha, Qatar, during the Christmas week of

2025, captained by Torstein Bae, pedagogically guiding us through the jungles of complexities towards Magnus Carlsen's magnificent wins. I am also indebted to Arne Tjølsen who volunteered assisting me in getting hold of the right data-files from FIDE.

References

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* London: Springer.

Gumbel, E. J. (1919). *Vier Jahre Lüge.* Berlin: Verlag Neues Vaterland.

Gumbel, E. J. (1922). *Vier Jahre Politischer Mord.* Berlin: Verlag der neuen Gesellschaft.

Gumbel, E. J. (1958). *Statistics of Extremes.* Cambridge: Columbia University Press.

Hjort, N. L. (2026). The stochastic view used in climate sciences: (some) perspectives from (some of) mathematical statistics. In *The Hasselmann Legacy: Stochastic Thinking in Climate Science*, L. Lin & H. von Storch, eds. Berlin: Springer Verlag, pp. xx–xx.

Hjort, N. L. & Stoltenberg, E. A. (2026). *Statistical Inference: 600 Exercises and 100 Stories.* Cambridge: Cambridge University Press.

Footnote. After having been absorbed by the most exciting chess in '10,000 Lifetimes' during the Christmas Week of 2025 I was inspired to dig out data, construct and test my models (along with a few discarded attempts, like the Pareto and the truncated normal, neither of which work well here). I then wished to write up a good *FocuStat Blog Post*, as these are sometimes reaching a sizeable international readership (we've got fanmail from Steven Pinker, etc.). To my frustration & irritation the University of Oslo has changed the Wondrously Well-Working Vortex `html`, which I and FocuStat comrades have used for ten years, for a Horribly Bad-Working Vortex editing system, where writing math is a nightmare; having the new Vortex system installed has also seriously disturbed some of our previously perfect pitch layouted Blog Posts. I might have to move to Andorra to find a university taking better care of its humble employees.