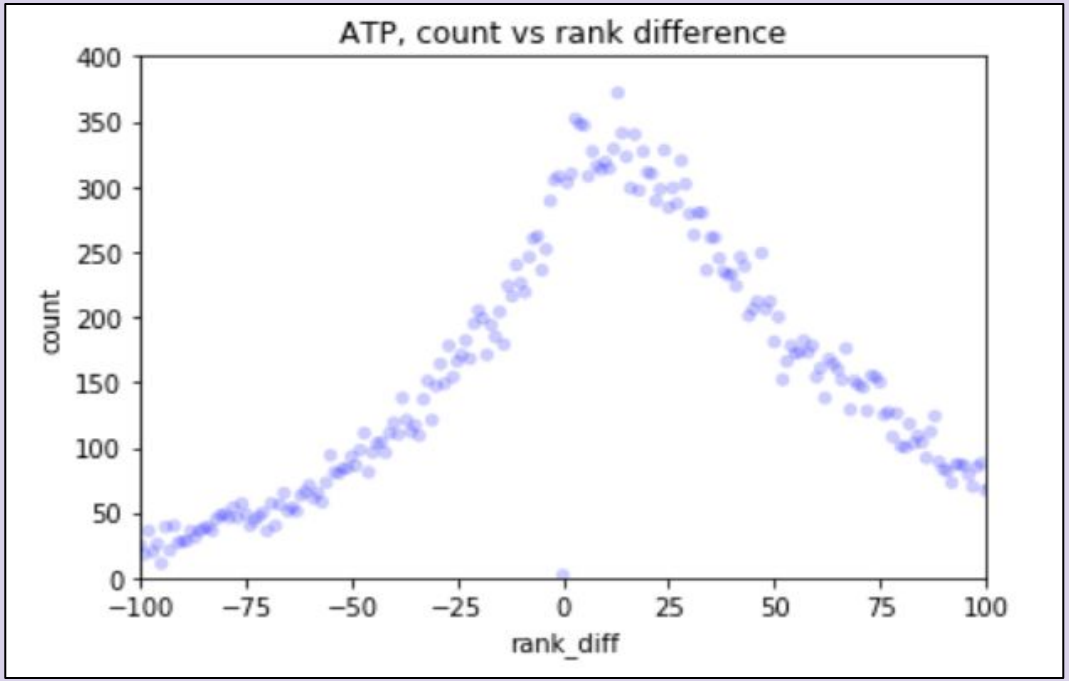
DSC 530 Term Project

Analysis of Upsets in Tennis

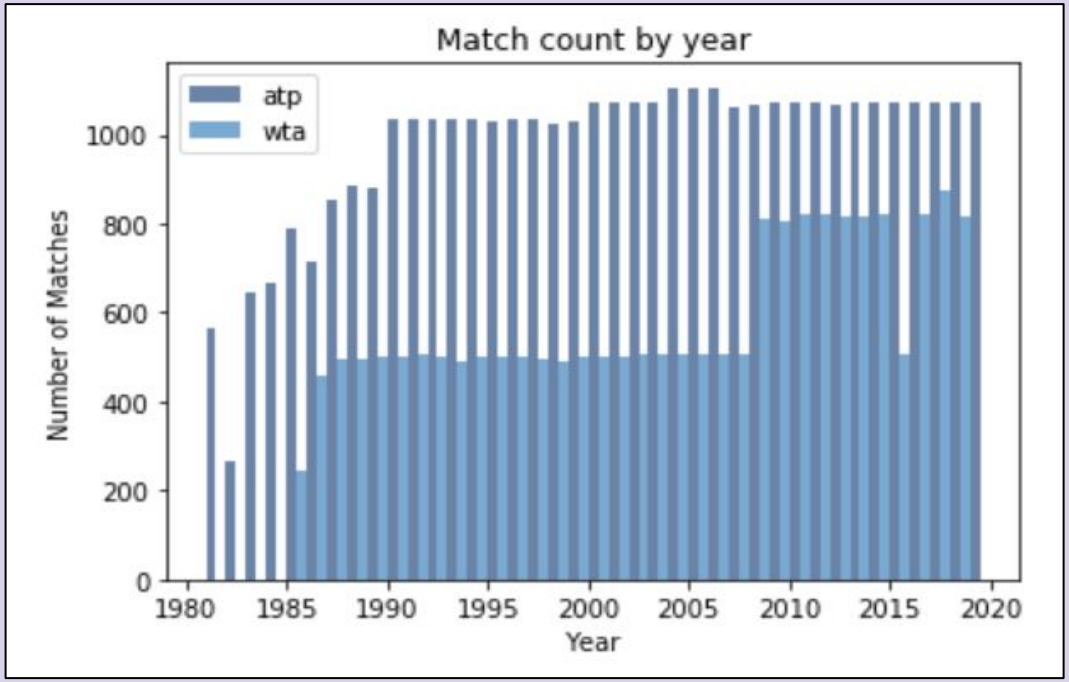
Christopher Briggs
May 30, 2020

	body	surface	tourney_level	tourney_date	round	winner_rank	loser_rank	rank_diff	upset	best_of
45709	wta	Hard	PM	20170306	F	15.0	8.0	-7.0	True	3
45710	wta	Hard	PM	20170306	SF	8.0	3.0	-5.0	True	3
45711	wta	Hard	PM	20170306	SF	15.0	26.0	11.0	False	3
45712	wta	Hard	PM	20170306	QF	3.0	7.0	4.0	False	3
45713	wta	Hard	PM	20170306	QF	8.0	21.0	13.0	False	3
45714	wta	Hard	PM	20170306	QF	26.0	14.0	-12.0	True	3
45715	wta	Hard	PM	20170306	QF	15.0	13.0	-2.0	True	3
45716	wta	Hard	PM	20170306	R16	3.0	16.0	13.0	False	3
45717	wta	Hard	PM	20170306	R16	7.0	10.0	3.0	False	3
45718	wta	Hard	PM	20170306	R16	21.0	5.0	-16.0	True	3

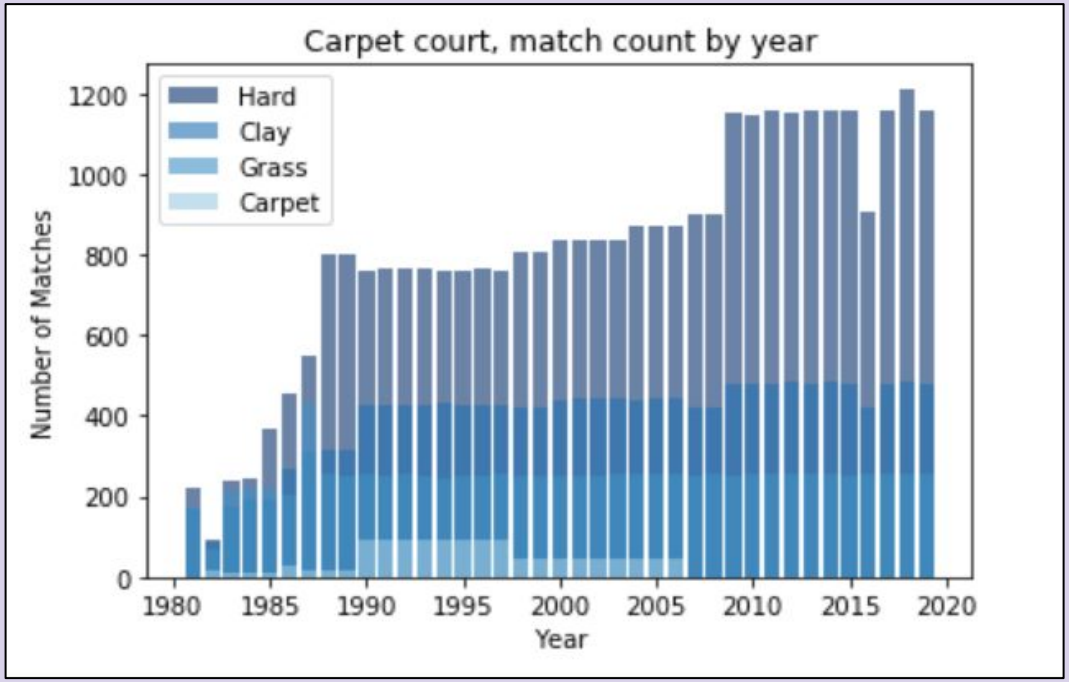
The variables used in the analysis are the governing body, surface type, tournament date, winner's rank, and loser's rank. Using these, I recode the rank difference, whether the result was an upset, and tournament format (i.e. best of 3 or 5 sets).



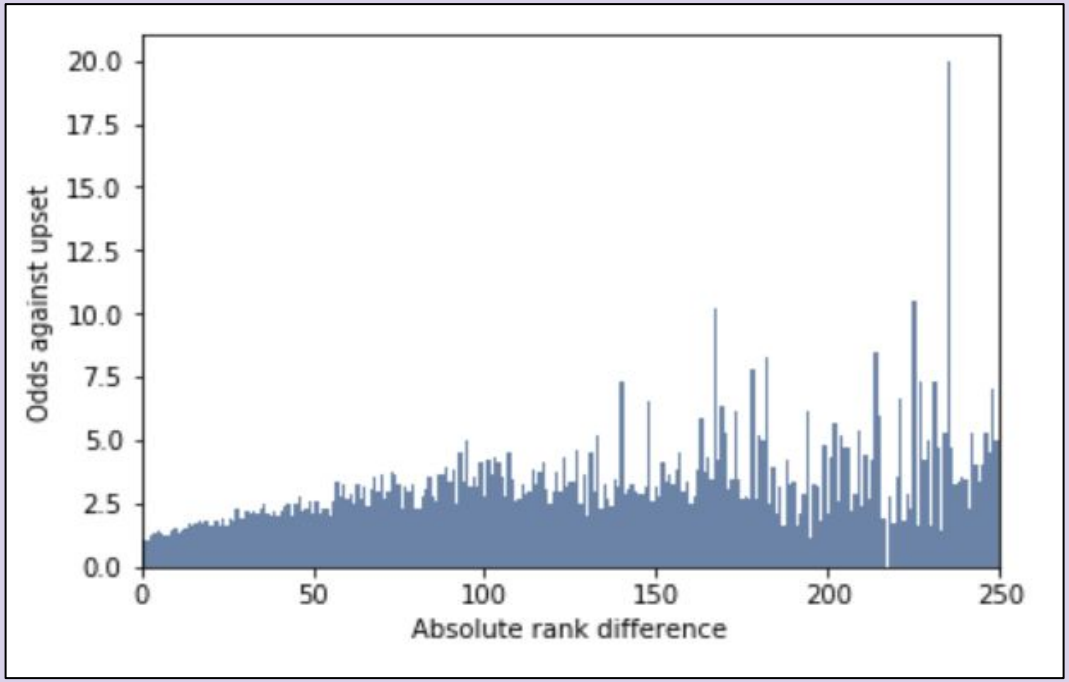
We see a slight rightward skew, which is expected, as the better player should on average win.



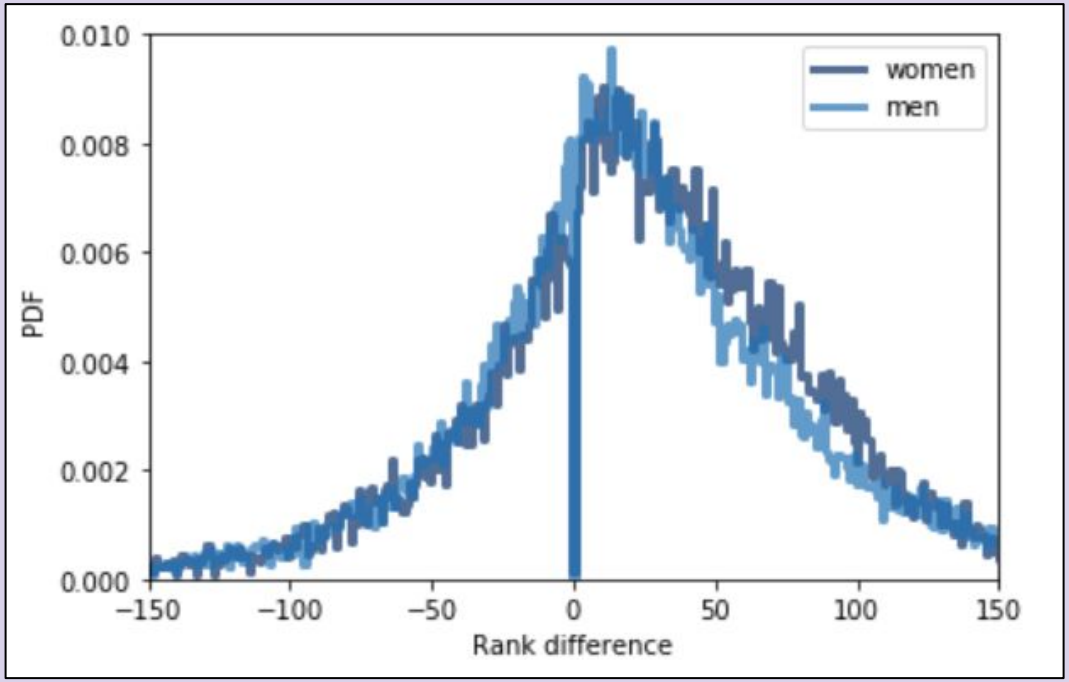
There is a break in 2009, as tournaments were reclassified for the women. This was also the year that men's Masters tournaments switched to a best of 3 sets format.



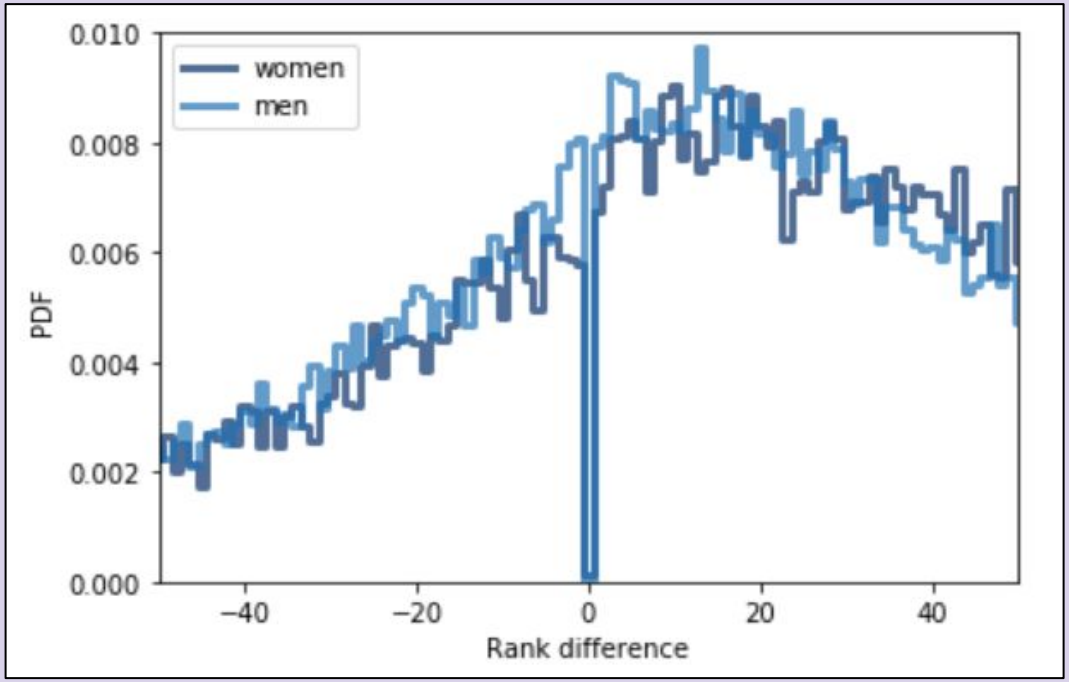
Carpet's loss is hard court's gain.



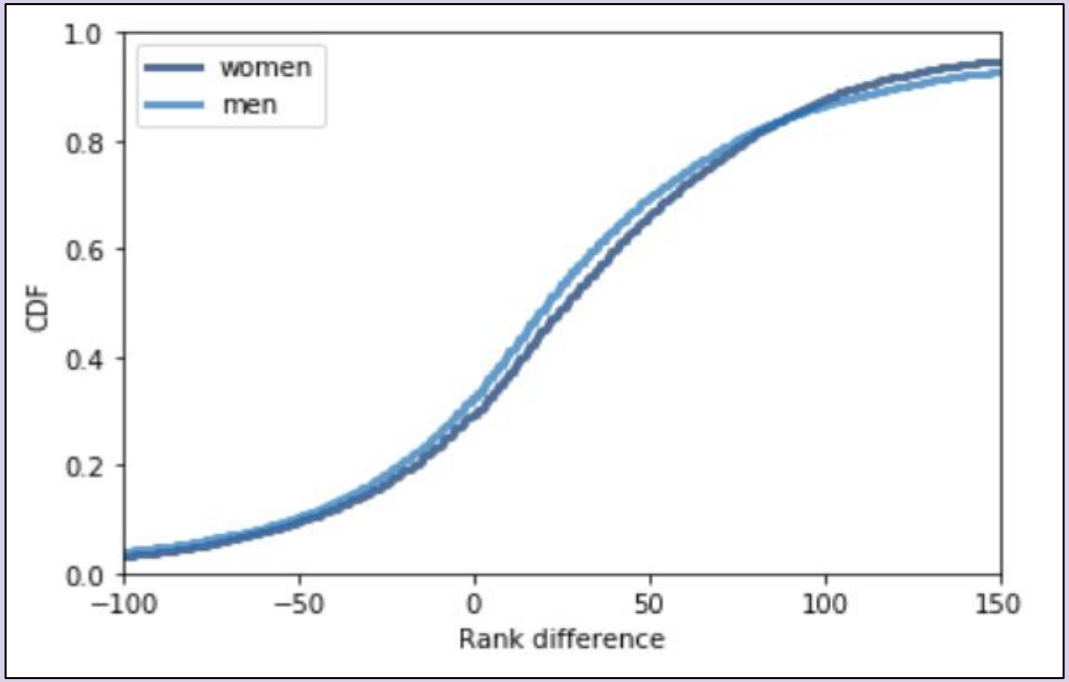
As one should expect, the odds of an upset decreases as the absolute difference in rank increases. The graph gets spikier as we move right because of diminishing quantities of data per ratio.



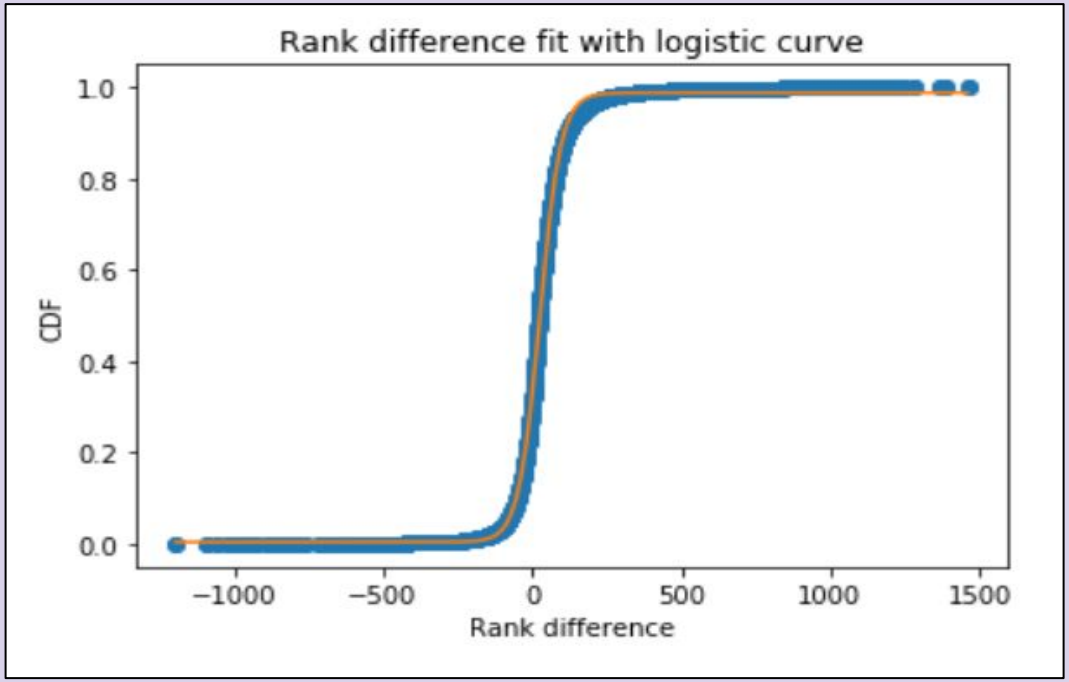
Both men and women are skewed rightward, as expected. It appears that the men's right tail is flatter than the women's.



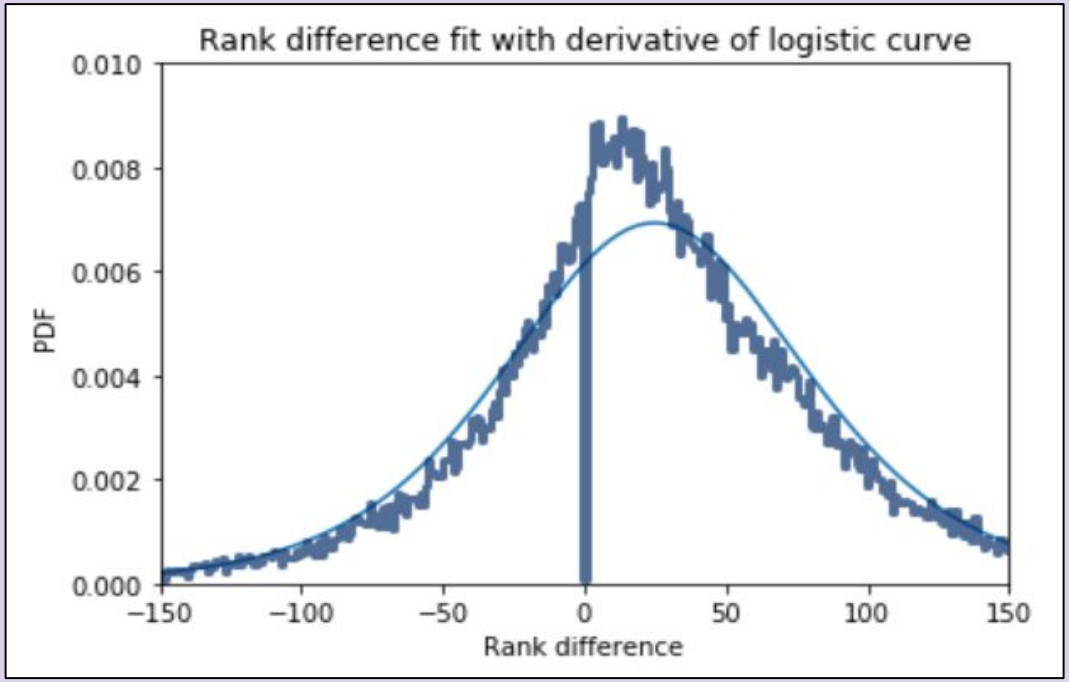
There is a clear outlier from the trend at 0. This is because by organizing body rules, two players cannot have the same rank. Should they have the same number of points there are tiebreaker rules.



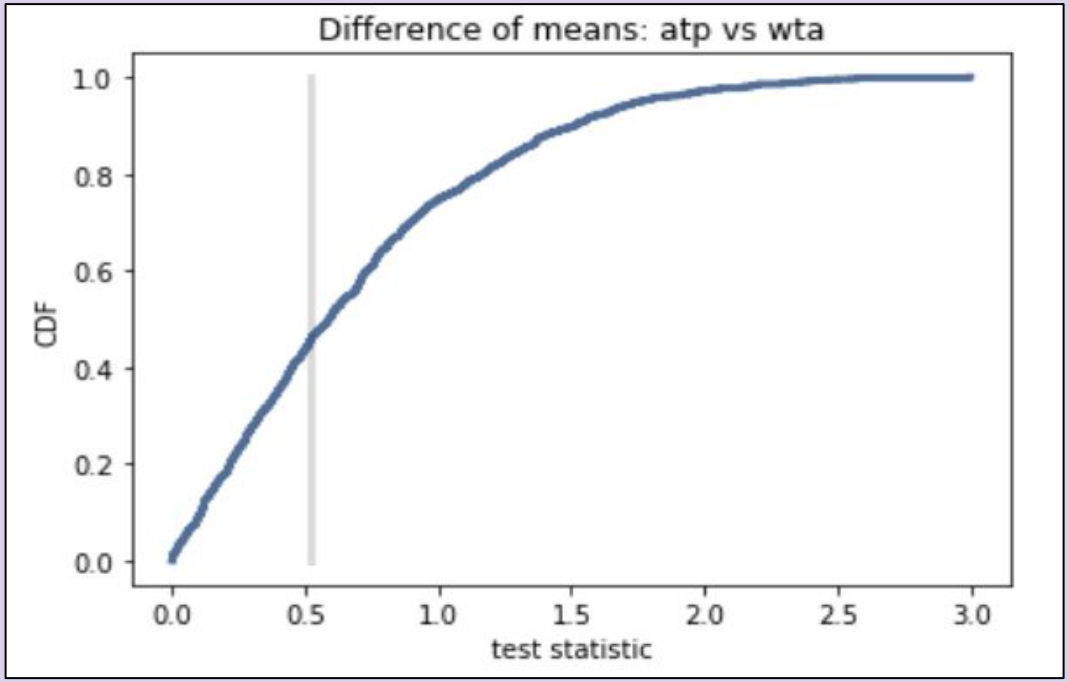
The relative elevation of the men's curve near the mode again shows what we might have missed in the PMF graph, which is that while the women's PMF is visibly fatter on the right tail, so is the men's tail a bit fatter on the left.



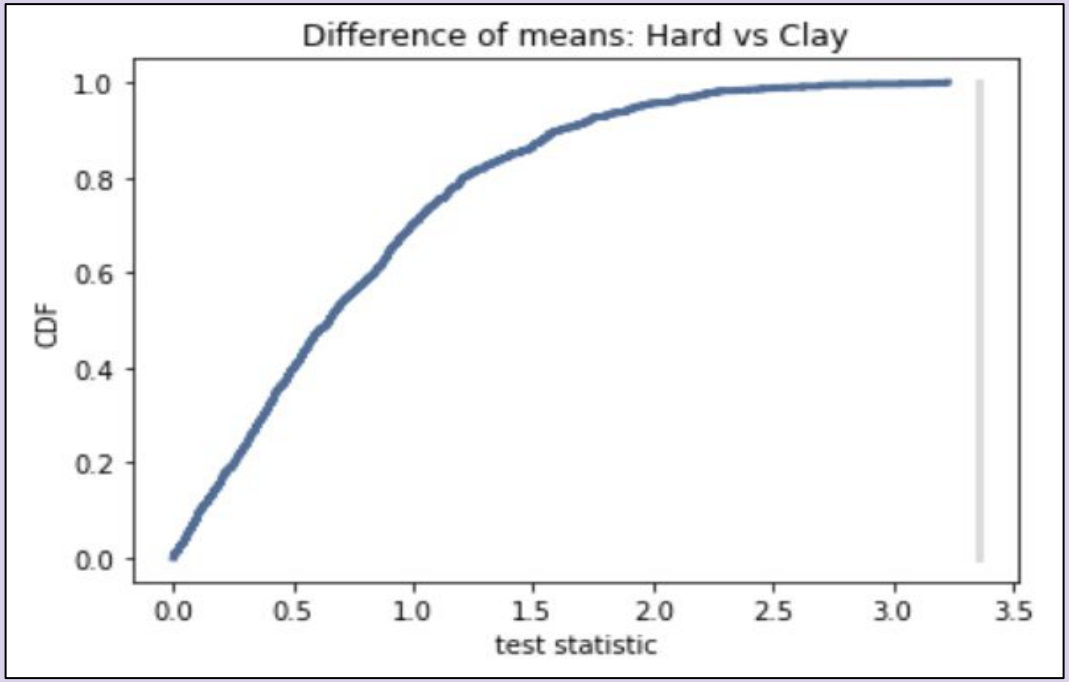
The logistic curve of best fit: $$f(x) = 0.988 - \frac{0.984}{1+e^{0.0282(x-24.8)}}$$



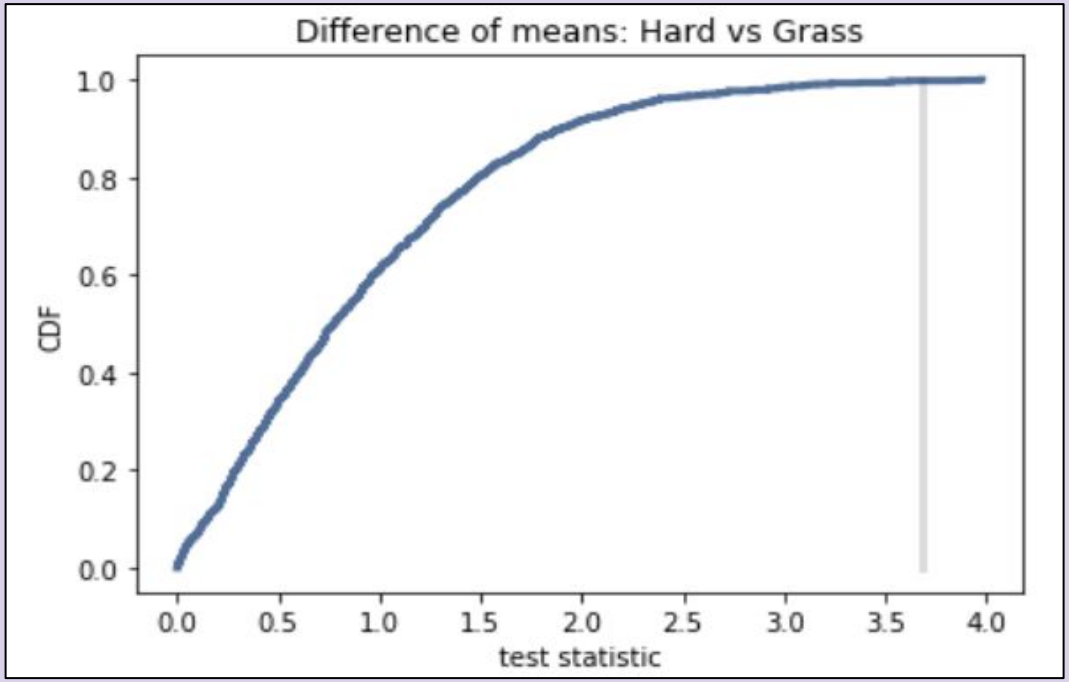
After differentiating, we can compare the generalized exponential analytic distribution of best fit to the PMF. The fit is alright but not excellent.



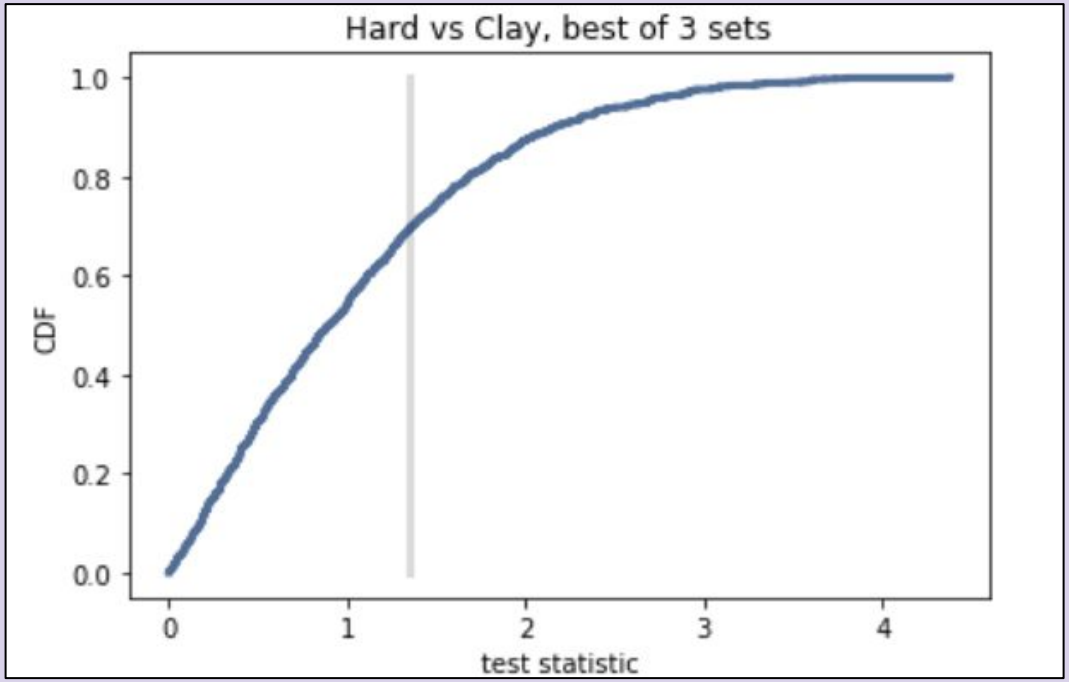
The hypothesis test of different means for the ATP and WTA is inconclusive. There is no substantial evidence that they have different means.



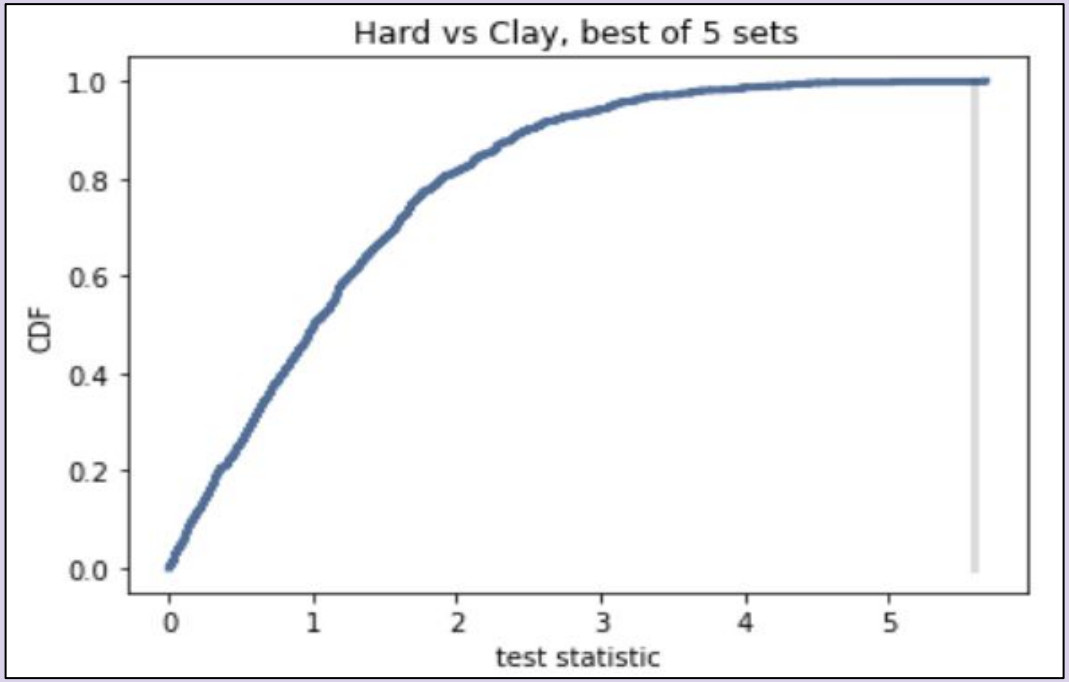
The hypothesis test of different means for hard courts vs clay is conclusive. We have good evidence that upsets are more likely on clay than hard courts.



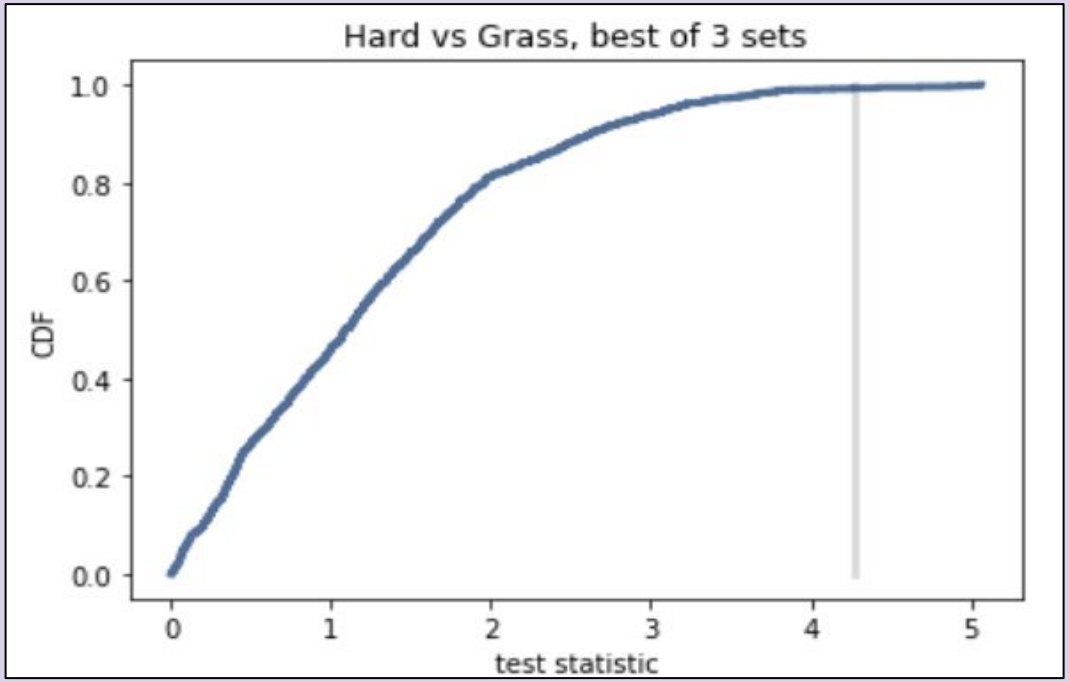
The hypothesis test of different means for hard courts vs grass is conclusive. We have good evidence that upsets are more likely on hard courts than grass.



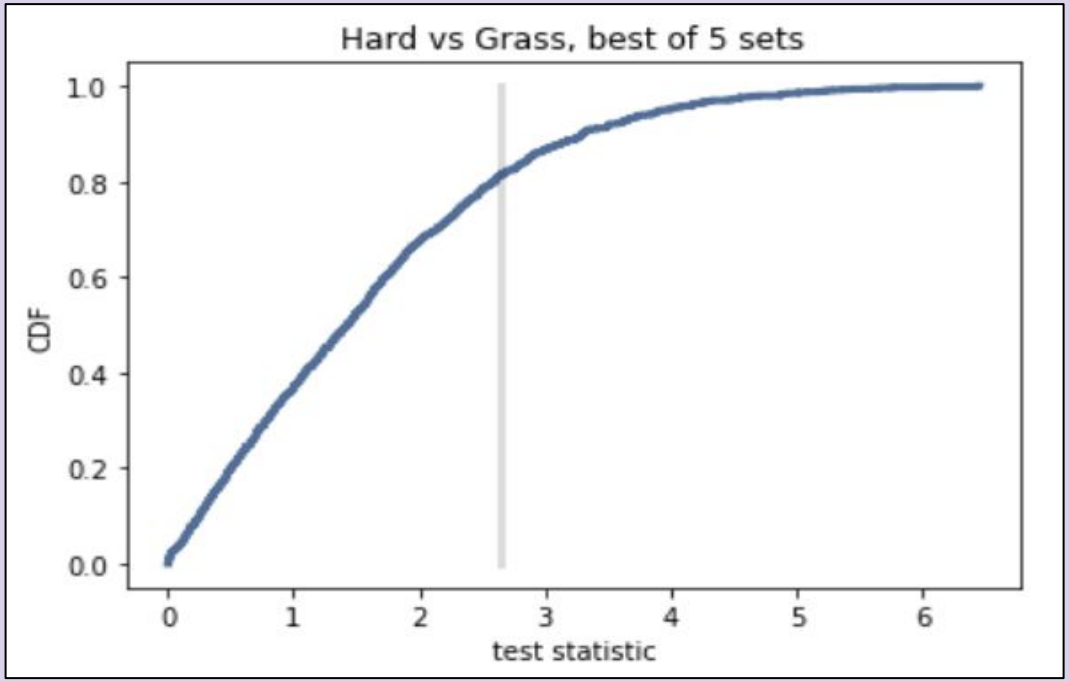
When playing best of 3 sets, there is no significant difference between means of rank difference on hard courts versus clay.



When playing best of 5 sets, there is a significant difference between means of rank difference on hard courts versus clay.



When playing best of 3 sets, there is a significant difference between means of rank difference on hard courts versus grass.



When playing best of 5 sets, there is no significant difference between means of rank difference on hard courts versus grass.