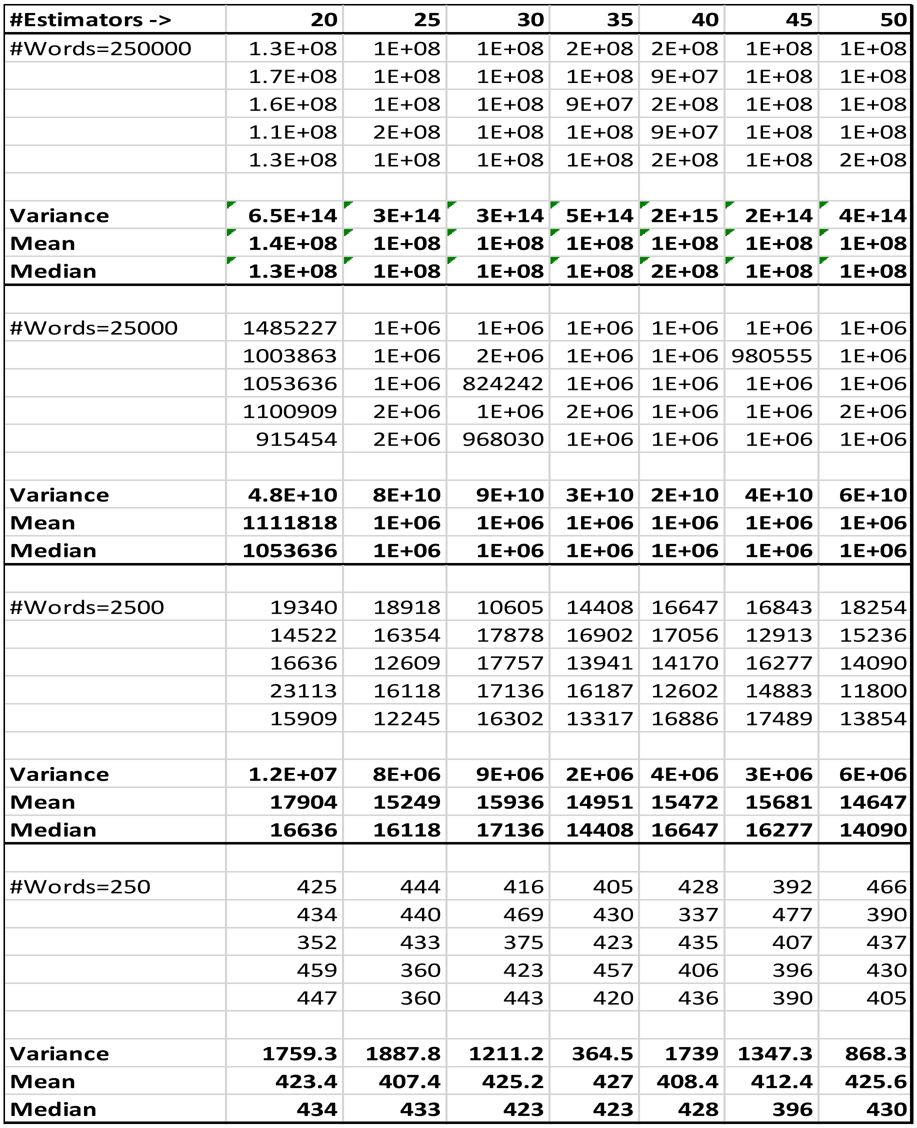
ESTIMATING THE IDEAL VALUE OF S

1. METHOD

The main function is put in a for loop to obtain 11 F2 estimates and the mean of those estimates, for each of the 250000, 25000, 2500 and 250 number of words.

A total of five trials is conducted for each of the above-mentioned number of words for different values of S: 20,25,30,35,40,45,50

2. TABLE



#Estimators ->	**20**	**25**	**30**	**35**	**40**	**45**	**50**
#Words=250000	1.3E+08	1E+08	1E+08	2E+08	2E+08	1E+08	1E+08
	1.7E+08	1E+08	1E+08	1E+08	9E+07	1E+08	1E+08
	1.6E+08	1E+08	1E+08	9E+07	2E+08	1E+08	1E+08
	1.1E+08	2E+08	1E+08	1E+08	9E+07	1E+08	1E+08
	1.3E+08	1E+08	1E+08	1E+08	2E+08	1E+08	2E+08
Variance	**6.5E+14**	**3E+14**	**3E+14**	**5E+14**	**2E+15**	**2E+14**	**4E+14**
Mean	**1.4E+08**	**1E+08**	**1E+08**	**1E+08**	**1E+08**	**1E+08**	**1E+08**
Median	**1.3E+08**	**1E+08**	**1E+08**	**1E+08**	**2E+08**	**1E+08**	**1E+08**
#Words=25000	1485227	1E+06	1E+06	1E+06	1E+06	1E+06	1E+06
	1003863	1E+06	2E+06	1E+06	1E+06	980555	1E+06
	1053636	1E+06	824242	1E+06	1E+06	1E+06	1E+06
	1100909	2E+06	1E+06	2E+06	1E+06	1E+06	2E+06
	915454	2E+06	968030	1E+06	1E+06	1E+06	1E+06
Variance	**4.8E+10**	**8E+10**	**9E+10**	**3E+10**	**2E+10**	**4E+10**	**6E+10**
Mean	**1111818**	**1E+06**	**1E+06**	**1E+06**	**1E+06**	**1E+06**	**1E+06**
Median	**1053636**	**1E+06**	**1E+06**	**1E+06**	**1E+06**	**1E+06**	**1E+06**
#Words=2500	19340	18918	10605	14408	16647	16843	18254
	14522	16354	17878	16902	17056	12913	15236
	16636	12609	17757	13941	14170	16277	14090
	23113	16118	17136	16187	12602	14883	11800
	15909	12245	16302	13317	16886	17489	13854
Variance	**1.2E+07**	**8E+06**	**9E+06**	**2E+06**	**4E+06**	**3E+06**	**6E+06**
Mean	**17904**	**15249**	**15936**	**14951**	**15472**	**15681**	**14647**
Median	**16636**	**16118**	**17136**	**14408**	**16647**	**16277**	**14090**
#Words=250	425	444	416	405	428	392	466
	434	440	469	430	337	477	390
	352	433	375	423	435	407	437
	459	360	423	457	406	396	430
	447	360	443	420	436	390	405
Variance	**1759.3**	**1887.8**	**1211.2**	**364.5**	**1739**	**1347.3**	**868.3**
Mean	**423.4**	**407.4**	**425.2**	**427**	**408.4**	**412.4**	**425.6**
Median	**434**	**433**	**423**	**423**	**428**	**396**	**430**

3. TABLE EXPLAINED

The table values are means of each of the 11 runs and 5 such trials for each of #Words = 250000, 25000,2500 and 250 and for each #Words, value of S=20,25,30,35,40,45,50.

The means, variances and medians are further computed for each of #Words = 250000, 25000,2500 and 250 and for each #Words S=20,25,30,35,40,45,50. So essentially we have means of means, medians of means and variances of means.

4. CHOOSING THE BEST VALUE FOR S-DECISION ANALYSIS

a) **Ranking Criteria**

Ranking #Estimators by mean, median and variance for each of the #Words. The closer the means and medians are to the Actual F2 in terms of absolute difference, the higher their rank. The lower the variance, the higher the rank.

b) **Choosing the 'Winner'**

For each of the #Words, the top three estimators are chosen for each of the mean, median and variance ranks. The estimators which are in the top 3 for at least 2 of the three ranking attributes are shortlisted as the likely candidates.

Finally, the candidate appearing for most #Words categories is chosen the 'winner'.

5. RANKINGS TABLE

	RankingMean	RankingMed	Ranking Variance	Winner
#Words=250000	40	35	45	20,50,40
	20	20	40	
	50	50	25	
	35	25	30	
	25	45	50	
	45	30	35	
	30	40	20	
#Words=25000	50	25	40	40,35
	35	40	35	
	40	45	20	
	25	50	45	
	45	35	50	
	30	20	25	
	20	30	30	
#Words=2500	30	40	35	40,30,45,20
	45	20	45	
	20	30	40	
	40	45	50	
	25	25	25	
	35	35	30	
	50	50	20	
#Words=250	35	50	35	50,35,30
	50	40	50	
	30	25	30	
	20	20	45	
	45	30	40	
	40	35	20	
	25	45	25	
			Final Winner	**40**

6 CONCLUSION

As the candidate 40 is shortlisted for #Words 250000, 25000 and 2500(3 out of 4), we choose it as the winner and the candidate for the best value of S.