How much to steal?? An experiment

- **Experiment: Divide 3M words of Wall Street Journal into 2 halves. Compare word counts across the two halves.**

Count 1	Count 2	Word	Count 1	Count 2	Word
1	0	abacuses	1	2	abilities
11	6	abandon	86	72	ability
29	21	abandoned	1	0	ability...
4	8	abandoning	0	1	ablaze
0	2	abandonment	192	149	able
2	0	abandons	0	1	able-bodied
1	0	abashed	4	9	abnormal
0	2	abate	0	2	abnormalities
1	1	abated	0	2	abnormality

Conclusion: *The smaller the count the worse the estimate* of what the count will be in a new, unseen data set.

Data from Mark Liberman

Key idea: *Deleted Estimation*

- **For all words with a given count in the first half, just use the average count in the 2nd half of those same words as the smoothed count**
- **Better: do both halves then average**
- **Key Idea:**
  - *Pool all items with the same frequency*
  - Compute average counts across halves
  - *Replace raw counts with the smoothed estimates*

Half 1	Half 2	Half 1	Half 2
0	1.60491	8	7.53499
1	0.639544	9	8.27005
2	1.59014	10	9.50197
3	2.55045	11	10.0348
4	3.49306	12	11.2292
5	4.45996	13	12.7391
6	5.23295	14	12.5298
7	6.28311	15	14.1646