Good Turing Smoothing

- Key Idea: **Assume total # unseen events to be n$_1$ (# of singletons), and adjust all the seen events in the same way**

Adjusted count

Sum of counts of all terms that occurred c(w,d)+1 times

$$p(w|d)=\frac{c^*(w,d)}{|d|};\ c^*(w,d)=\frac{c(w,d)+1}{n_{c(w,d)}}n_{c(w,d)+1}, 0^*=\frac{n_1}{n_0}, 1^*=\frac{2*n_2}{n_1}, \ldots..$$

Share the counts among all the words that occurred c(w,d) times

$$n_r = the\ \ number\ \ of\ words\ \ with\ \ count\ \ r$$

$$What\ \ if\ \ n_{c(w,d)}=0?\ \ What\ \ about\ \ p(w|REF)?$$

Heuristics are needed