The Population Frequencies of Species and the Estimation of Population Parameters

By I. J. Good

Biometrika, Vol. 40, No. 3/4.
(Dec., 1953), pp. 237-264.

Presented by Eugene Weinstein

Problem Statement

- Start with an infinite population of animals
- Sample *N* animals from the population
- Want to estimate the population frequency: how often a species occurs in the population
- Problem: sample may not be representative

Why do we care about animals?

- In general, want to estimate occurrence frequency based on a corpus of data, e.g.
- Given *x* hours of speech
  - See how often word B is heard after word A
  - Want to estimate how often A→B happens in spoken dialogue
- Primary application: language modeling

Estimating Population Frequencies

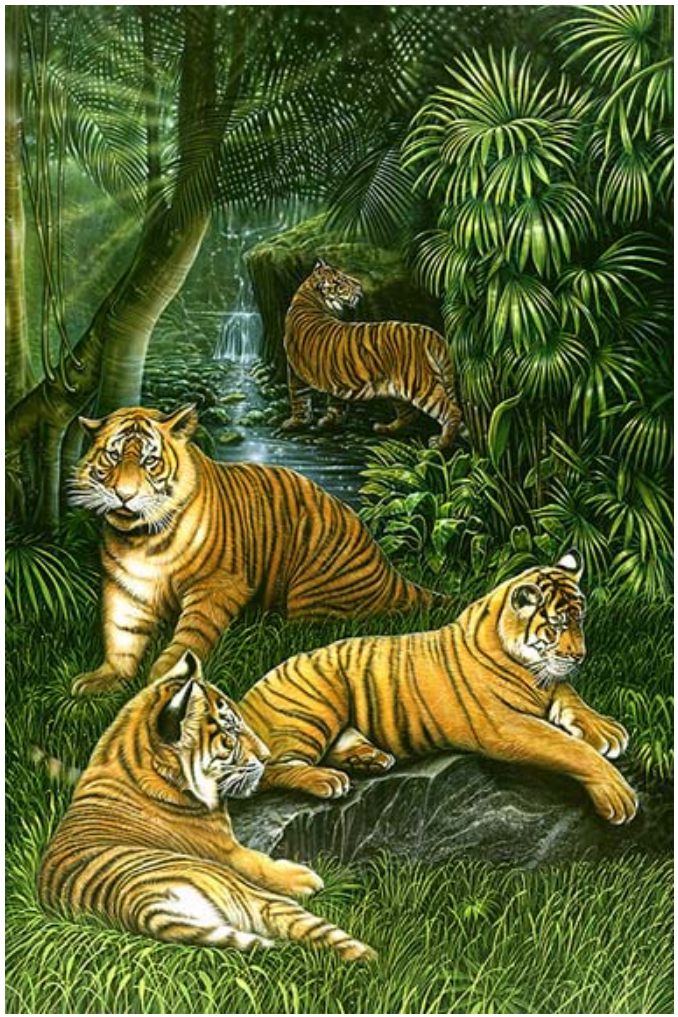
- Sample *N* animals from the population
- Say a species occurs *r* times in sample
- Intuitive estimate: "*r/N* of population belongs to this species"
- Main problem: Species that did not occur in sample are not included in estimate
- Another problem: when *r is* small, this is a bad estimate

Estimating Population Frequencies

- Sample *N* animals from the population
- Let n_r be the number of species occurring *r* times in the sample, such that

$$\sum_{r=1}^{\infty} r n_r = N$$

- Want to estimate q_r : Population frequency of species occurring *r* times in the sample



An Example: Words in newspaper English (N=43,989; S=6,001 unique words)

r	n_r
1	2976
2	1079
3	516
4	294
5	212
6	151
7	105
8	84
9	86
10	45

Population Frequency Estimate

- Good-Turing Estimate of expected population frequency

$$\mathrm{E}[q_r] = \frac{r^*}{N} \qquad r^* \approx (r+1)\frac{n_{r+1}}{n_r}$$

- Why Turing? It was his idea!

- Caveat: we need to first "smooth" the values $n_r \rightarrow n_r'$

$$r^* = (r+1)\frac{n'_{r+1}}{n'_r}$$

Smoothing

- Smoothing the observed frequency counts makes our approximation “good”
- No definitive smoothing approach is given
- But, some suggestions are made
  - e.g., approximate data by freehand sketch
- And, some statistical guidelines are given

Total Frequency Estimate

- Expected total frequency of all species in the sample is

$$1 - \frac{n_1}{N}$$

- Meaning the probability that the next animal sampled will belong to a species unseen in the original sample is

$$\frac{n_1}{N}$$

Total Frequency Estimate Derivation

- Number of species occurring *r* times in the sample: n_r
- Expected total frequency of all species that occur *r* times in the sample: $n_r \bullet \mathrm{E}[q_r] \approx (r+1)\, n_{r+1} / N$
- So the total frequency of all species occurring *r* times or more ≈

$$\sum_{m=r}^{\infty} n_m \bullet \mathrm{E}[q_m] = \frac{(r+1)n_{r+1} + (r+2)n_{r+2} + \cdots}{N}$$

- Total frequency of all species at all in the sample

$$\sum_{m=1}^{\infty} n_m \bullet \mathrm{E}[q_m] = \frac{2n_2 + 3n_3 + \cdots}{N} = \frac{N - n_1}{N} = 1 - \frac{n_1}{N}$$

An Example: Words in newspaper English (N=43,989; S=6,001 unique words)

r	n_r	n_r'	r^*	$E[q_r]_= r^*/N$
1	2976	2961	0.73	0.0000166
2	1079	1075	1.4	0.0000318
3	516	509	2.4	0.0000546
4	294	305	3.4	0.0000773
5	212	209	4.4	0.0001000
6	151	153	5.4	0.0001228
7	105	118	6.2	0.0001409
8	84	91	–	–
9	86	70	–	–
10	45	–	–	–

- Meaning: A word that was seen twice in our corpus should be seen every 1/0.0000318 = 31,420 words

An Example: Words in newspaper English (N=43,989; S=6,001 unique words)

- A foreigner learns English from this corpus
- How long should it be before she sees a word she hasn't learned?
- n_1/N = 2976/43,989 = 0.067
- She will encounter a new word every 6.7% of words read
- Say she learns only S-n_1= 3025 words
- Then, new word every $(2n_2+n_1)/N \approx$ 11.6% of words read

Population Parameters

- Paper gives estimates of population parameters measuring heterogeneity
- e.g., entropy

Another Example

- From: KW Church and WA Gale (CSL 1991)
- Training sample: ½ of the 1998 AP wire
- N=22,000,000 bigrams (word A → word B)
- Assume English has V=400,000 words
- Primary difference: we know total population size: $V^2 = 1.6 \cdot 10^{11}$
- Thus, can directly calculate r^* when $r=0$

An Example: Bigrams in 1998 AP Wire (N=22,000,000)

r	n_r	r^*
0	74,671,100,000	0.0000270
1	2,018,046	0.446
2	449,721	1.26
3	188,933	2.24
4	105,668	3.24
5	68,379	4.22
6	48,190	5.19
7	35,709	6.21
8	27,710	7.24
9	22,280	8.25

- Main point: Assign 0.0000270 probability to unseen bigrams
- Seen bigrams must “pay the price” with probability