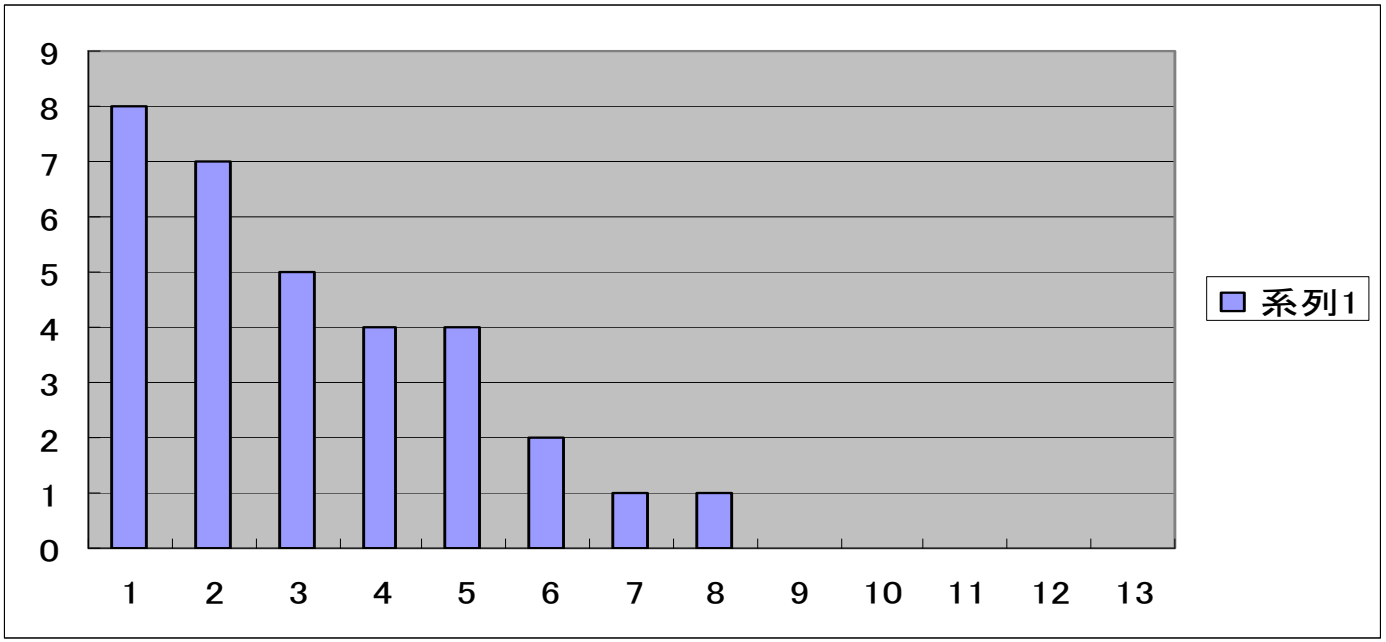
スムージング

未出現事象の扱い

- 観測データ(教師データ)の数が十分に大きくない場合は、本来、出現する可能性のある事象が教師データに含まれないことがある。
  - 例:サイコロを5回振って教師データを作っても、出ない目が必ず一つはある。
  - 例:新聞記事10年分のテキストから単語の出現頻度を計算しても、出現していない単語があるかもしれない。
- 本来、出現する可能性がある事象ではあるが、観測データに出現していないものの真の生起確率をどのように評価しておけばよいか?→スムージング
  - これは、未知のデータを扱う場合に重要

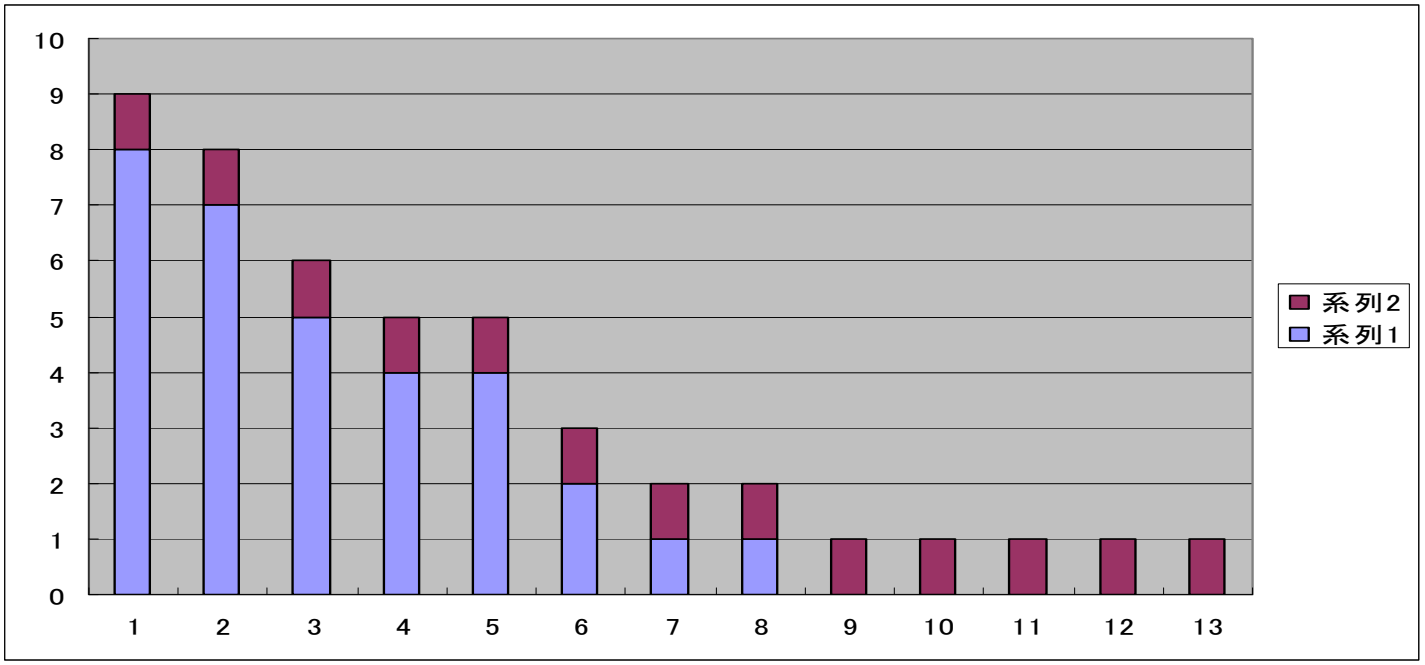
Back-off smoothing（元データの頻度）



実際に出現した単語(8個)

出現していないが、これから出現する可能性がある単語(5個)

各単語の頻度にδ(＝1)を加算

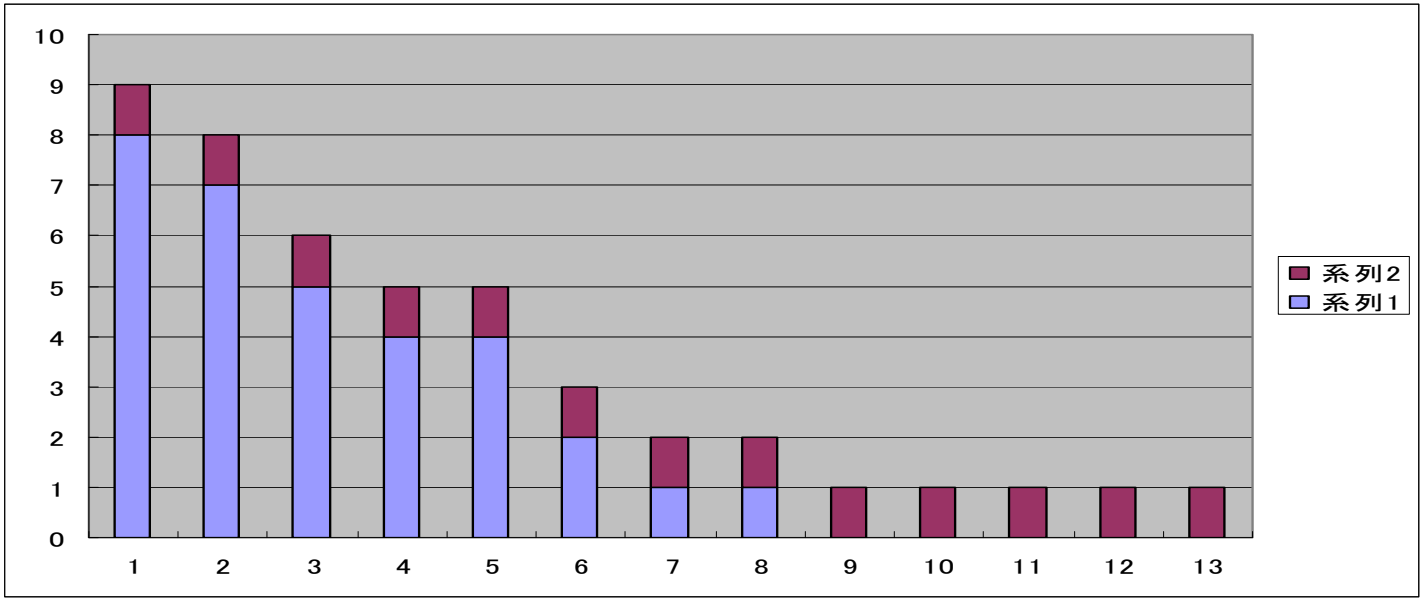


実際に出現した単語(8個)

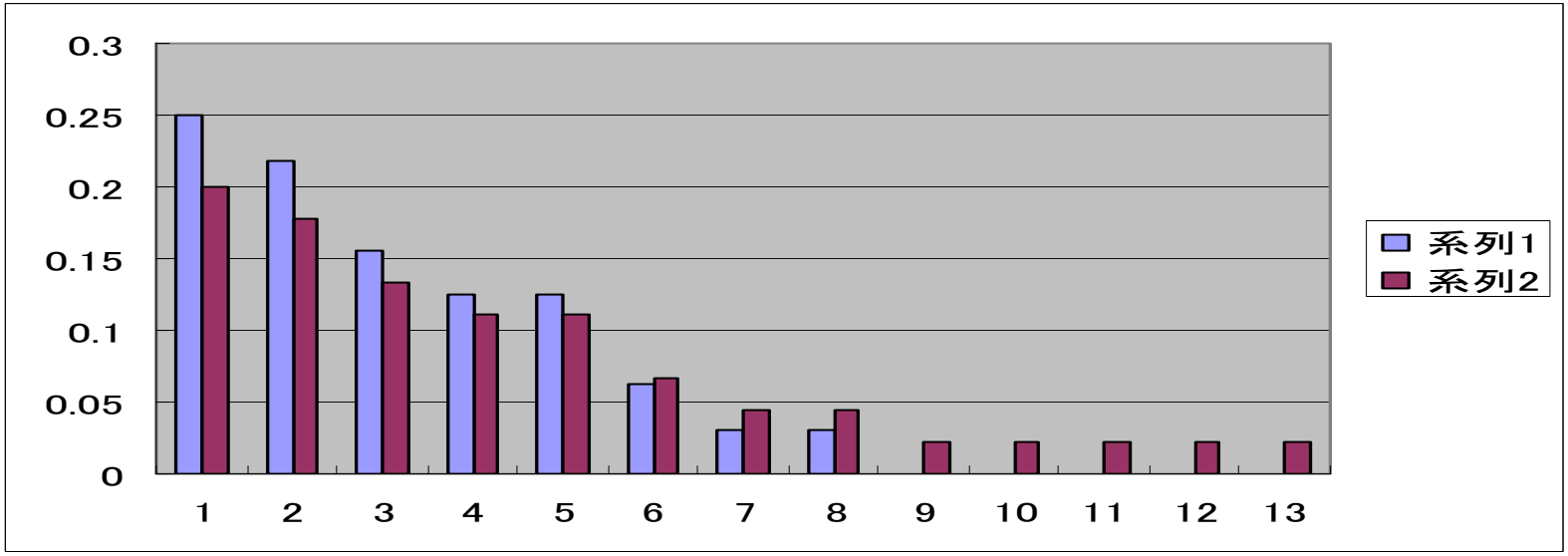
出現していないが、これから出現する可能性がある単語(5個)

Back-off smoothing（確率を計算しなおす）

原データ



確率



単語の生起確率を求める場合のスムージング
Good-Turingの推定

- Good-Turingの推定

  語数Nのコーパス中でr回出現する異なり単語数をn_rとする。すると

$$N = \sum_{r>0} r \times n_r = n_1 + 2n_2 + 3n_3 + \ldots$$

  ここでコーパスにr回出現する単語wの頻度を次の式で推定するのがGood-Turingの推定

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

  注）rが最大のときはr+1が定義できないので、r*=r とするしかない。r が小さいときが重要なので、これでもよいであろう。

- Good-Turingの推定

  語数Nのコーパス中でr回出現する単語の数をn_rとする。すると

$$N = \sum_{r>0} r \times n_r = n_1 + 2n_2 + 3n_3 + \ldots$$

ここでコーパスにr回出現する単語wの頻度を次の式で推定するのがGood-Turingの推定

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

- ここで0回出現した単語の出現頻度の期待値0^*は

$$0^* = \frac{n_1}{n_0} = \frac{n_1}{\text{全語彙数} - \text{コーパスに出現した語彙数}}$$

- 一方、1回以上出現した単語の相対頻度の総和を求めると

$$\sum_{r>0} \frac{n_r \times r^*}{N} = 1 - \frac{n_1}{N}$$

つまり、 $\frac{n_1}{N}$ がコーパスに出現しない全単語の頻度の合計の推定確率

- なお、 $d = \frac{r^*}{r}$ をディスカウント係数という。

Good-Turingの推定の導出

- 母集団における異なり単語数をMとする
- 母集団における単語w_iの出現確率を$P(w_i)$
- w_i が語数（サイズ）Nのコーパス中で出現する回数を$C(w_i)$ 当然 $\sum_{i=1}^{M} C(w_i) = N$
- 単語wがコーパス中にr回出現したとき、wの母集団での生起確率および出現回数の期待値は

$$E[P(w) \mid C(w) = r] = \sum_{i=1}^{M} P(w = w_i \mid C(w) = r) P(w_i) \qquad -(1)$$

$$r^* = E[r \mid C(w) = r] = E[P(w) \mid C(w) = r] N \qquad -(2)$$

- サイズNのコーパスにおける単語の出現確率分布を2項分布とすると

$$P(w=w_i \mid C(w)=r)=\frac{P(C(w_i)=r)}{\sum_{i=1}^{M} P(C(w_i)=r)}$$

$$=\frac{{}_N C_r P(w_i)^r (1-P(w_i))^{N-r}}{\sum_{i=1}^{M} {}_N C_r P(w_i)^r (1-P(w_i))^{N-r}} \quad -(3)$$

この結果を(1)に代入すると

$$E[P(w) \mid C(w)=r]=\frac{\sum_{i=1}^{M} {}_N C_r P(w_i)^{r+1} (1-P(w_i))^{N-r}}{\sum_{i=1}^{M} {}_N C_r P(w_i)^r (1-P(w_i))^{N-r}} \quad -(4)$$

サイズNのコーパス中にr回出現する単語の総数の期待値

$$E_N[N_r] = \sum_{i=1}^{M} P(C(w_i) = r) = \sum_{i=1}^{M} {}_N C_r P(w_i)^r (1 - P(w_i))^{N-r}$$

すると(4)は${}_N C_r = \frac{r+1}{N+1} {}_{N+1} C_{r+1}$から以下のように書き換えられる

$$E[P(w) \mid C(w) = r] = \frac{r+1}{N+1} \frac{E_{N+1}(N_{r+1})}{E_N(N_r)} \qquad -(5)$$

この結果を使って(2)式のr*を求めると

$$r^* = E[P(w) \mid C(w) = r]N = N \frac{r+1}{N+1} \frac{E_{N+1}(N_{r+1})}{E_N(N_r)} \qquad -(6)$$

ここでNが十分大きく、$E_N(N_r)$をコーパス中に出現頻度N_rで近似すると

$$r^* = (r+1)\frac{N_{r+1}}{N_r} \quad \text{となる} \qquad 0^* = \frac{N_1}{N_0}$$