Analyzing Linguistic Data　　2008/11/06　野村真理子

3　Probability distributions

3.1　Distributions（分布）

確率変数の確率分布は、異なる結果の可能性を示す。

確率変数は２つの重要なカテゴリーに分類される。

DISCRETE：整数の値をもつ不連続のもの　（例）頻度数のような確率変数

CONTINUOUS：実数の値をもつ連続的なもの　（例）継続的測定のような確率変数

3.2　Discrete distributions (離散型分布)（pp.44-48）

＊Table 3.1

CELEX lexical database (1860 万語のコーパスにおける英語の単語の頻度リスト)に基づく 4 つの語－高頻度語 *the*, 中頻度語 *president*, 低頻度語 *hare, harpsichord*－の頻度と相対頻度（語の頻度をコーパスサイズで割ったもので、英語におけるこれらの語の確率の推定）

＊Urn mode（壷モデル）

テキスト生成の最も単純なモデルでは、テキストに含む語の選択は壷から石を抽出することに似ていて、赤い石を抽出する確率は、壷の中の赤い石の割合によって得られる。

復元抽出し、時が経っても語の確率は変化しないと仮定する。

独立性（1 回目の試行の結果が 2 回目の試行の結果に影響を及ぼさない）を仮定する。

→これらの仮定はかなり単純化したもの

実際の言語において、*the the* の確率は非常に小さい。

＊Table 3.2

Table 3.1 の 4 語の確率（相対頻度）、Brown corpus における観測度数と期待度数

期待度数の計算方法

（例）the: 1000000＊0.05885575=58856 tokens

Brown corpus の総語数＊CELEX における the の相対頻度

the と *president* の期待度数は観測度数より小さく、*hare* はより大きく、*harpsichord* はちょうど同じ

＊語の頻度の分布の特性についていくつか仮定をすることが必要

Brown corpus では president は 382 回出現しているが、もし同じサンプリングの基準で同じ種類のテキストからさらにコーパスが構築されたら、president の出現数はコーパスにより異なる。→コーパスのある語の頻度は、確率変数である。

100 万語を抽出する繰り返される実験で、確率変数 *president* は Brown corpus の観測度数の 382 と似た値になると期待するが、本当に知りたいのは、コーパス間での president の頻度の変動の規模である。

＊*p*: PROBABILITY OF SUCCESS と *q*: PROBABILITY OF FAILURE

p: ある特定の語を観測する確率　*q*: 他の語を観測する確率

q=1－*p*　（例）hare　*p*=0.0000082, *q*=0.999991

NUMBER OF TRIALS (n): コーパスサイズ

語の頻度：パラメータ *p* と *n* を持つ BINOMINALLY DISTRIBUTED RANDOM VARIABLE

（2 項分布の確率変数）

2 項分布の特性は良く知られていて、それによりコーパス間の語の頻度の変動がどれだけかよりよく洞察できる。

＊区別する必要がある 2 種類の特性：POPULATION の特性と SAMPLE の特性

母集団の特性：無限の実験で平均して起こると期待するものを考える

標本の特性：限られた普通は小数回の実験で実際に起こったことを考える。

（例）*president* に対して *p*=0.000133(CELEX による), *n*=1,000,000 であれば、観測度数 382 は驚くべきものであるかどうか知りたい→これは母集団の問題。100 万語の無数の標本でこの頻度を何回観測するか？、この頻度は平均して期待するものに近いか？

この本では、ほとんどの場合母集団の特性を用いるが、ときにはある特定のサイズの標本がどのようなものであるか知ることも役立つ。R には両方に対するツールがある。

＊Figure 3.1

上のグラフ：*the, hare, harpsichord* の 3 語が出現すると期待される頻度の確率を示したもの

サイズ *n*=1,000,000, 語の population probability *p*（CELEX における相対頻度による）で関数 **dbinom()**を用いて確率を計算

(FREQUENCY FUNCTION, PROBABILITY DENSITY FUNCTION)

左上（the）：横軸が頻度、縦軸が頻度の確率

パラメータ *n*=1,000,000, *p*=0.059 で 2 項分布していると仮定

```
> dbinom(59000, 1000000, 0.05885575)
        頻度        n              p
[1] 0.001403392
```

＊*the* のグラフ：Table 3.2 の期待度数の 58856 に集中している頻度を示している。

60000 より大きい値を観測する確率はごく小さいので、Brown corpus で観測された 69971 の頻度に驚くには確固とした理由がある。

hare のグラフ：それぞれの頻度に対してそれぞれ高密度の線、対称的でない。

最も高い確率は 0.1391 で頻度 8 に対して起こり、Table 3.2 の期待値と一致

Brown corpus の観測値 1 は、明らかに低い。

harpsichord のグラフ：頻度 0 の方が Table 3.2 の頻度 1 よりもわずかに可能性が大きい。

＊下のグラフ：100 万語のコーパス 500 のみ（cf. 上のグラフは無数）で観測する頻度の確率

分布の形は不規則（左のグラフの不規則性ははっきり見える、真ん中でもある程度）