Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

金田　拓

3 Probability distributions

✓ 3.2 Discrete distributions　pp. 48~53　(in pp. 44~58)

➢ ある単語が求める語である確率を *p*、試行回数（コーパスサイズ）を *n* としたとき、

- 確率＝*p*の分布に対する影響を図示したのがFigure3.1. (p. 47)
- 試行回数＝コーパスサイズ*n*の分布に対する影響を図時したのがFigure3.2.(p. 49)

確率が低くても試行回数が少なくても、分布の対照性が失われるのが見てとれる。

➢ R で 2 項分布に対して用いることのできるツールは 4 種類。

1. dbinom(x, n, p)：確率密度。x 回成功する確率を求める。*n* は試行数、*p* は成功確率
2. qbinom(q, n, p)：変位値。q%までで最も多い成功回数を求める。
3. pbinom(x, n, p)：累積分布。x 回以下成功する確率を求める。0~x 回成功する確率の累積。
4. rbinom(k, n, p)：乱数。乱数を生成してシミュレーションを行い、合計 k 回観測する。

➢ CELEX の中で、ある単語が hare である確率は 0.0000082 だったが、100 万語のサンプルの中で hare が 1 回見つかる確率を dbinom( )関数を使用して求めてみると、以下のコマンドとなる

```
> dbinom(1, size = 1000000, prob = 0.0000082)
```

size=, prob=の部分は省略して、値だけ入力しても計算可能。

➢ 同じく hare が 0~1 回見つかる確率を求める場合、3 通りの計算方法がある

```
> dbinom(0, size = 1000000, prob = 0.0000082) + dbinom(1, size = 1000000, prob = 0.0000082)
> sum(dbinom(0:1, size = 1000000, prob = 0.0000082))
> pbinom(1, size = 1000000, prob = 0.0000082)
```

上段は個別に頻度を計算してから足す方法。中段はベクトルを使って求める場合。下段は成功数が 1 以下である確率を求める場合（成功数はマイナスにはならないので 0 が下限）

➢ 100 万語の Brown コーパスでは、単語 president は 382 回出現した。CELEX での出現確率をもとに、president が 100 万語で 381 回以下出現する確率を計算してみる。

```
> 1 - pbinom(381, size = 1000000, prob = 0.00013288)
```

すると、確率は 0 であるという答えが返ってくる。

➢ サンプル 1000 語の中に単語 the の見つかる数を図式化する

```
> n = 1000
> p = 0.05885575
> frequencies = seq(25, 95, by = 1)                #25, 26, 27, ..., 94, 95
> probabilities = dbinom(frequencies, n, p)
> plot(frequencies, probabilities, type = "h", xlab = "frequency", ylab = "probability of frequency")
```

コマンドは実質 4 行目から。2 項分布を仮定して、成功確率約 0.059 の試行を 1000 回したとき、25~95 回成功する確率を足して図式化する(1000 語の中に the が 25~95 回出現するのは何件かをプロットする)。成功すると Figure3.2 (p. 49)の左の図となる。

1. 成功確率 *p* で *n* 回試行するのを *s* 回シミュレーション観測した場合（単語 hare が 100 万語の中に何回現れるかを、500 回観測する）を求めるには、まず値を代入する

```
> s = 500
> n = 1000000
> p = 0.0000082
```

2. 次に結果をクロスタブに集計して、何回成功したケースが何件観測されたのか（hare が何回出現したケースが 500 回のうち何件あったか）求める。その後観測数 *s* で割って、割合に変換する。（シミュレーションなので計算するたび異なった値が現れる）

```
> x = xtabs( ~ rbinom(s, n, p) ) / s
> x
```

3. x の中身が確実に数として扱われるよう as.numeric( )関数を用いて、上のクロスタブを図式化してみる。

```
> plot(as.numeric(names(x)), x, type = "h", xlim = c(0, 30),
+ xlab = "frequency", ylab = "sample probability of frequency")
```

➢ 他 2 つの 2 項分布に使用できる関数

・ 成功確率 0.5 の試行を 10 回やって､4 回以下成功する確率を累積して計算するコマンド。

```
> pbinom(4, size = 10, prob = 0.5)
```

・ 成功確率 0.5 の試行を 10 回やって、成功 0 回から累積して 0.3769531 までの確率で現れる成功数の最大値。0.3769531 の確率で成功する数ではなく、累積値であることに留意。

```
> qbinom(0.3769531, size = 10, prob = 0.5)
```

- オランダ語の定冠詞”het”が、小説 1000 語に何回出現しているかを求めてみる

- オランダ語の定冠詞"het"の現れた回数のテーブルを確認する

```
> havelaar$Frequency
```

- 定冠詞“het”の頻度は 2 項分布に従って分布しているといえるか、ということを検証するために、2 項分布に従った場合と実数値とを重ね合わせてプロットして、視覚的に判断する。*n* は試行の回数、*p* は 1 サンプル（小説の 1000 語）あたり het が何回現れたか、その平均。

```
> n = 1000
> p = mean(havelaar$Frequency / n)
```

- 2 項分布の図に重ねてみて判断するための下準備として、0.005~0.995 まで、0.01 刻みのベクトルを qnts に格納

```
> qnts = seq(0.005, 0.995, by=0.01)
```

1. まずは 2 項分布を仮定して、小説内での"het"の平均出現頻度 *p* からして、*n*=1000 回の試行中 qnts の確率で成功するのは何回以下かを連続的に plot する。
2. 次に実際の"het"回数と、何回以下成功するかの確率を連続的にプロットする。

```
> plot(qbinom(qnts, n, p), quantile(havelaar$Frequency, qnts), xlab = paste("quantiles of (", n, ",", round(p, 4), ")-binomial", sep=""), ylab="frequencies")
```

☆ コマンドが成功すると Figure3.3 の図となる。実数値は直線ではなく、特に高頻度で現れる確率になるほど、2 項分布を仮定した値から著しくずれていることがはっきりと確認できる。