Analyzing Linguistic Data 2008/10/23 野村真理子

2.2 Visualizing single random variables (pp. 21-24)

＊棒グラフとヒストグラム：１つの確率変数の分布の visual summary

data set (ratings) 動植物を表す 81 語に対して収集された数種類の rating を持つデータ
各語に３つの rating: weight, size, subjective familiarity

> colnames(ratings) Class: 因子（動物か植物か）
変数：さまざまな言語特性（Frequency, Length, SynsetCount, etc.）

Figure 2.1 左上：語の長さの bar plot 関数 barplot( )で作成される

> barplot(xtabs(~ratings$Length), xlab="word length", col="grey")

xlab: X 軸のラベル col: bar の色を grey に設定

※５文字が最も多く、分布はやや非対称的

関数：平均 mean( ), 中央値 median( ), 範囲 range( ), 最小値 min( ), 最大値 max( )

> mean(ratings$Length)
> median(ratings$Length)
> range(ratings$Length)
> min(ratings$Length)
> max(ratings$Length)

＊ヒストグラムの作成 MASS package の中の関数 truehist( )で作成

準備：MASS package をインストール→パッケージを読み込む > library(MASS)
（パッケージを閉じる > detach(package:MASS)）

Figure 2.1

> truehist(ratings$Length, xlab= "word length", col="grey")
> truehist(ratings$Frequency, xlab= "log word frequency", col="grey")
> truehist(ratings$SynsetCount, xlab= "log synset count", col="grey")
> truehist(ratings$FamilySize, xlab= "log family size", col="grey")
> truehist(ratings$DerivEntropy, xlab= "derivational entropy", col="grey")

※ヒストグラムでは、bar の縦軸の面積の合計が１になる。たとえば、右上のヒストグラムから、5 文字と 6 文字の長さの語が合わせてデータの 40%以上を占めることが分かる。

※一番下のグラフは、非常に非対称的な分布。このデータのほとんどの語は、形態的 family member を持っていない。

＊いくつかのグラフを１つの window に表示する方法

Rのデフォルトは、１つの window に１つのグラフであるので、デフォルトを関数 par ( ) を使って変更する。 > par(mfrow = c(3, 2)) （行, 列）を指定 ３行２列で並べられる

デフォルトに戻す方法： > par(mfrow = c(1, 1))

par( )：他の多くのパラメータ（色、フォントサイズ、マージンなど）を設定できる（後で順次紹介される） on-line help が利用可能： ?par or help(par)