2013 年度 森林統計学 第 2 回 データの記述 配布資料-教科書の補足

1. “(計測)単位”と“精度”(p.13、2 番目の段落)、”有効数字”について

教科書の例では、「ある体重が例えば 152 ポンドと記録されたとすれば, 真の値は 151.5 ポンドと 152.5 ポンドの間にあるとみなされる.」となっている（より正確には、151.5000...ポンドと 152.4999....ポンドの間にあると考えられる; 1 ポンド(lb) = 0.45359... kgf)。この場合、体重計の計測単位（計測精度）は 1 ポンドなので、真の体重は 151.7 ポンドかもしれないし、152.1 ポンドかもしれない。計測精度より細かいことはわからないため、152 ポンドと表記するのである。

他の例として、カエルの足の長さが「8cm」と計測されたとする。1cm の精度の計測ならば「8cm」で正しいが、1mm（=0.1cm）の精度で計測されたならば「8.0cm」と表記しなければならない。前者の場合有効数字は 1 桁、後者の場合有効数字は 2 桁となる。

※データを表記するときに小数点以下の桁をそろえることも重要で、例えば 0.1cm の精度で計測された場合、「6.5cm, 7.0cm, 8.1cm, 3.5cm, 5.0cm, ... 」等となる。これを「6.5cm, 7cm, 8.1cm, 3.5cm, 5cm, ... 」などとしてしまうと、0.1cm の計測精度のデータと 1cm の計測精度のデータが混在していることになってしまう（よくある間違いなので注意)。

有効数字について：水の量 1ℓは 1000ml とも表現できるが、これでは計測精度は不明である。10ml の精度で計測したとしたら、1.00ℓあるいは 1.00×10^3ml と表現するべき（10ml のコップ 100 杯分; “100”の 3 桁が有効桁）で、このように表現すれば有効数字はℓで表しても ml で表しても 3 桁となる。実学では有効数字と計測精度を常に意識する必要がある。

なお、計測単位は 1,000（千円単位での家賃など)、5 や 0.5（5mm 単位の測量など）の場合もある。また、例えば木の数（本数）を計測する場合、本数は離散型変数で計測単位は通常は 1 だが、「立木の ha あたり本数」などの場合には、計測精度によって 1,000 本/ha や 100 本/ha が適切な計測単位になることもあるので注意すべし。

2. データの分類(p.14, 6～7 行目「データを分類するとき, 階級の数は 10 個ないし 20 個にするのがよいといわれている. 」)

ここでの「データの分類」とは、度数分布表を作る際の適切な階級数のことである。教科書では 10 あるいは 20 個がよいとされているが、実のところ一般的にはそうとは言えない。データの数が少ないと、10 階級ではデータがどのように分布しているかを適切に表すには分類が細かすぎることがあるし、逆にデータの数が多い時には 20 階級でも粗すぎる場合もある。

現在の統計学では、データ総数 n に応じて階級数 k を決めるのが良いとされている。その方法にはいくつかあるが、以下に 2 つの例を示す。

2-1. 永田（1992）* の方法

$$k \cong \sqrt{n} \tag{1}$$

2-2. スタージェスの公式 ** (Sturges' rule)

$$k \cong 1+\log_2 n = 1+\frac{\log_{10} n}{\log_{10} 2} \tag{2}$$

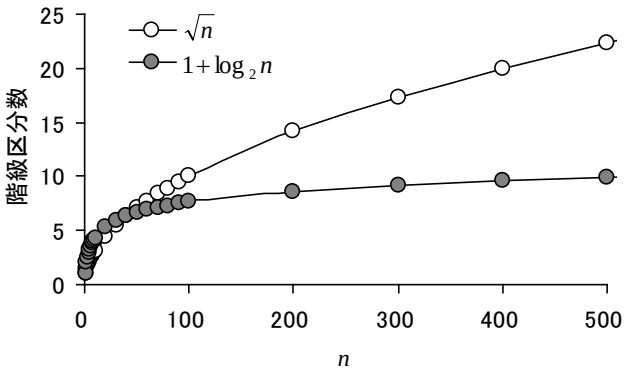
ここで、”≅” は「ほぼ」あるいは「おおむね」を意味し、右辺の式で得られた値を参考に k を決めるべしということである（”≒”も同じ)。例えば右辺の結果が 5.43 ... だった場合、四捨五入して $k=5$ としてもよいし、都合により $k=6$ としてもよい（最終的には分析者の判断による)。

---

* 永田靖（1992）度数表とヒストグラム.（入門統計解析法. 永田靖, 日科技連). 9～15.

** 中井検裕（1991）一次元のデータ.（統計学入門. 東京大学教養部統計学教室編, 東京大学出版会). 17～40.

n に対する(1), (2)式の結果は右図のようになる。$n \leqq 50$ 程度では両式の結果はあまり変わらない。スタージェスの公式はデータが二項分布（あるいは正規分布）に従うときには適切だが、一般には階級数は過少になるという指摘もある。いずれにせよ、(1)式、(2)式で得られる階級数は参考値である。最終的な度数分布表の階級数は、データの性質や度数分布表を使う目的を考慮して、分析者が判断することになる。



3. 度数分布表とヒストグラムの作り方

度数分布表は、単なる数値の羅列であるデータを、値がどのように分布しているかを視覚的に判断しやすくするために作成するヒストグラムの、もととなる表である。従って、その作成に際しての基本的な考え方は以下の通りである。

1）計測単位を明確にして、データを採取・整理する。
2）適切な階級数「データの分類」の数）を決める。
3) 階級幅を決める（決めた階級数で最小値から最大値までのデータをもれなく集計できるよう）。
※階級の境界ちょうどに重なるデータがないように工夫が必要（どちらの階級に含めるか曖昧になるため）
4）各階級に含まれるデータの数を数え上げる。

参考) 度数分布表とヒストグラムの作り方（永田, 1992*, を一部改変）

o 度数分布表（略して「度数表」ともいう）の作成手順

1. データの測定単位 m（測定の最小きざみ）を明確にして、サンプリングを行う。データ数を n とする。
2. データより最大値 x_{max}、最小値 x_{min}, 範囲 R を求める。(※ $R = x_{max} - x_{min}$)
3. 仮の区間の数 k を(1)式あるいは(2)式をもとにして決める。
4. 区間の幅 c を、$c \cong \frac{R}{k}$ をもとにして決める。この際、c は測定単位 m の整数倍に丸める。

   c をいくつにするかの最終的な判断は、度数分布表の作成者にゆだねられている。R/k が、まとめようとするデータの性質から考えて半端で不自然な値だと考えられる時には、適切な値に丸めるなどしてよい。逆に、c を R/k より細かくする判断も場合によってはありうる。ただし、集計するデータが階級の境界値と重ならないようにするために、c は必ず測定単位 m の整数倍である必要がある。
5. 各区間の境界値を次のように決める。

   一番下の境界値を、境界値とデータの値が一致しないようにするために、$x_{\min} - \frac{m}{2}$ とする。これに区間の幅 c を順次加えていって各区間の境界値とし、x_{max} を含むまで繰り返す(データの取りこぼしがないように)。
6. 階級値（区間の中心値；その区間の上側と下側の境界値の平均）を求める。
7. 階級値ごとにその階級（区間）に入るデータの個数（度数）を数え上げ、表にまとめる。

o ヒストグラムの作成手順

1. 度数表の階級値（区間の中心値）を横軸にとり、度数を縦軸に取る。
2. 度数表に基づいて、それぞれの区間の度数に応じて柱を描く。柱の幅は階級幅に比例させる（階級幅が同一の時はどの柱も同じ幅で描く)。柱の間は空けないのが慣例だが、離散型変数で階級幅が著しく異なる場合などには柱の間を空ける場合もある。

[3. 平均を示す線を書き込み、ヒストグラムの余白に n, $\bar{x}$, s(標準偏差; 不偏分散 s^2 から求めた教科書 p.21, (9)式によるものを用いるのが一般的）を記入する。]←(3.は次回の[課題 2]で作成する)