

2021年3月

区間推定の考え方【第49回生物統計学】

1 導入

統計関連の話題の中で「信頼区間」なる言葉を見かける機会が多いと思われます。指標がどのような意味を持つのか見方を知ることによって解析結果から得られる解釈の幅が大きく広がります。言葉の意味合いを間違えて捉えているケースも多いので正しく理解しましょう。

2 標本抽出によるブレを考慮する区間推定

母集団から抽出した標本から得られた平均値が推定される母平均と一致することは少ないというのは容易に想像できると思われます。標本の取り出し方に確率的なブレが生じるためです。したがってピンポイントで母平均を推定するより、母平均がおおよそどのくらいの範囲に収まる値なのか区間推定することが求められます。

母平均 θ を中心に $\theta \pm 40$ の範囲をとる一様分布を例に、標本平均が75であった場合、そこから θ の95%信頼区間を推定する方法を解説します。

区間の長さ80のうち、両末端から5%を削ると $40 * 0.025 = 1$ となるので想定できる標本平均の範囲は $\theta - 39 \leq X (=75) \leq \theta + 39$ となります。

これを、標本平均を代入して θ について解くと $34 \leq \theta \leq 114$ となります。

この区間[34, 114]が母平均 θ にとっての95%信頼区間であると見なすことができます。

信頼区間について「○○%の確率でこの範囲内に入る」と認識している方も多いと思いますが、母集団を定義するパラメータである θ について確率分布を考えるのは不適切です。「信頼係数○○%で繰り返し区間推定を行った場合、 θ が推定された区間の中に収まっている確率が○○%」というのが正しい解釈です。

面積の求め方の分かりやすい一様分布で例えましたが、母集団が他の分布を取っていても考え方は変わりません。

3 母比率の区間推定

区間推定は連続量について求めているものを見かける機会が多いと思いますが、二項分布を正規分布で近似することで比率についても信頼区間を推定することができます。

母集団から抽出したサンプルが特定の属性Aを持つ確率を p とするとき、 n 個を抽出した中で属性Aを持つサンプル数 X は二項分布 $\text{Bin}(n, p)$ に従います。



X の平均、分散は $E[X]=np$, $V[X]=np(1-p)$

標本中に占める属性 A もちサンプルの比率 X/n の平均、分散は

$E[X/n]= E[X]/n=p$, $V[X/n]= V[X]/n^2=p(1-p)/n$ となります。

サンプル数が十分に大きいとき二項分布は正規分布で近似できるため、比率 X/n は正規分布 $N(p,p(1-p)/n)$ に従います。

ここから正規母集団の母平均の区間推定と同様に母比率の信頼区間を推定できます。 X/n の 95%信頼区

間は $\left[\frac{X}{n} - 1.96 * \sqrt{\frac{p(1-p)}{n}}, \frac{X}{n} + 1.96 * \sqrt{\frac{p(1-p)}{n}} \right]$ になります。

また、群間の比率の差の信頼区間を求めることも可能です。

$$\left[\left(\frac{X_1}{n_1} - \frac{X_2}{n_2} \right) - 1.96 * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \left(\frac{X_1}{n_1} - \frac{X_2}{n_2} \right) + 1.96 * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right]$$

群間の比率の差の信頼区間が 0 を跨がない場合、差が 0 になる確率が極めて低いということであり、特定の属性に該当するサンプルの比率の差を比較するカイ 2 乗検定などで有意差が検出できる可能性が高いといえます。