



農場で使える統計講座 ～ 第6回 点推定その2 ～

かやの

先月は、「推定」とはなんなのかということについて説明しました。牛群全体や地域全体といった多くの対象牛について知りたい場合、すべてのウシについて調べることはほぼ不可能です。そのため、そうした対象集団（母集団）からサンプル（標本）を抽出して、それらを調べることで母集団を明らかにしよう！というのが推定でした。

点推定のバラツキ：不偏分散

前回（農場で使える統計講座 第5回目）の例では、乳量について平均値（標本平均）を出しました。繰り返しになりますが、母集団の平均（母平均）はわからないので（全部調べられないので）、標本から計算した標本平均をつかって、「集団の平均乳量はこれくらいである！」とってしまうのが推定です。

今回考えたいのは、**推定値のバラツキ**についてです。第2回目でバラツキを数値化するという事で、「**分散**」や「**標準偏差**」という指標について説明しました。A農場とB農場で平均乳量は同じくらいだけれど、分散や標準偏差が違うので、ウシの乳量に関してA農場とB農場でバラツキ具合が異なる（分布が違う）という話を覚えているでしょうか。もう一度分散の式を下に出します。

$$\text{分散} = \frac{1}{\text{データの数}} \sum_{i=1}^{\text{データの数}} (i\text{番目のデータ} - \text{平均})^2$$

実は、上の式で求める分散は正確には「標本分散」といい、母集団が全部把握できているときはコチラでOKです（牛群が10頭しかないA農場、B農場の例など）。しかし、実際には全頭分のデータがわかっていることはほとんどなく、抽出した標本を使うことになります。この場合、少し式が変わります。

$$\text{不偏分散} = \frac{1}{\text{データの数} - 1} \sum_{i=1}^{\text{データの数}} (i\text{番目のデータ} - \text{平均})^2$$

本来得たい母分散を偏りなく推定するという事で、「**不偏分散**」と呼ばれます。詳しいことは書きませんが、データの数が小さいと、通常の分散（標本分散）では母分散と同じにならず、母分散よりも小さくなるという性質があります。そのため、「データの数－1」を分母にすることで推定値が少し大きくなるように調整してやるイメージです。

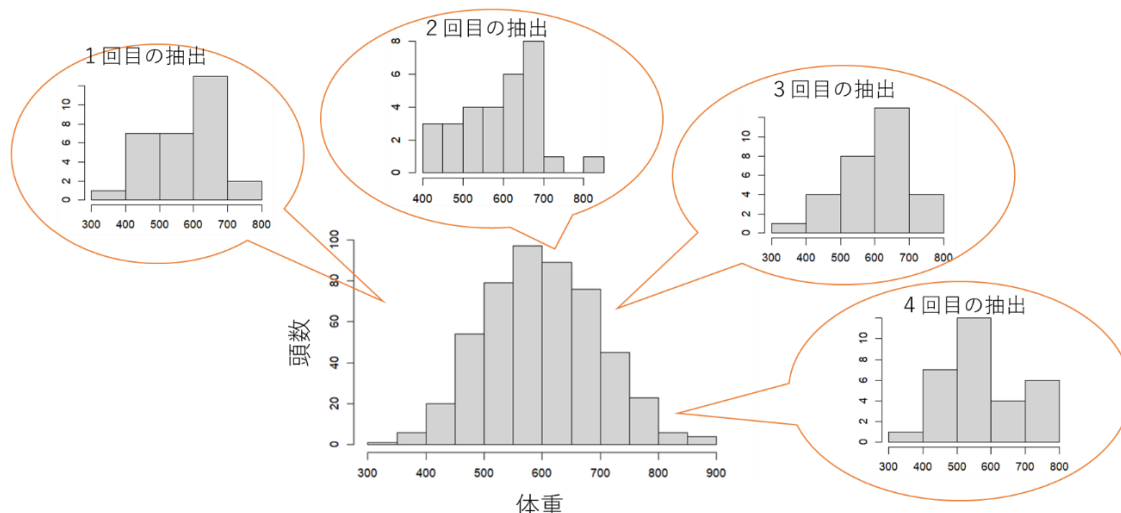
点推定のバラツキ：標準誤差

さて、ここからが今回のキモです。分散は二乗をしているので意味として直感的に理解しづらく、そのためルート（√）をとって単位をわかりやすくする、それが**標準偏差**であるという話を以前にしました。ですので、ここでも同じようにルートをとって…としたいですね。そうすると、不偏分散のルートは**不偏標準偏差**といいます。実は、これは直接的にはあまり使われません。ここでも詳しい話はスキップさせていただきますが、簡単にいうと、「注目したいのは標本そのもののバラツキよりも、平均値のバラツキだよ」ということになるからです。ということで、標本を使って乳量の平均をだしましたが、その「平均値」のバラツキを表すのが下に示した**標準誤差**という指標です。

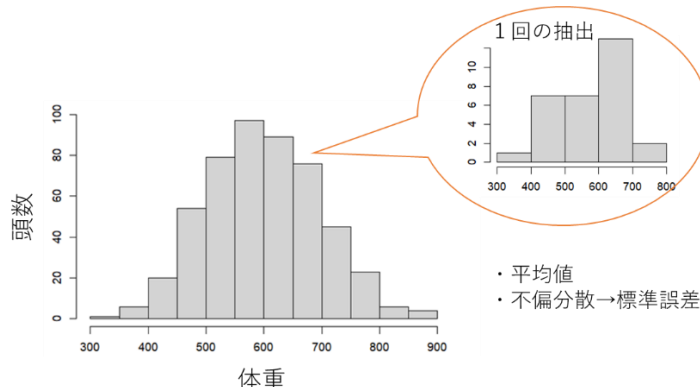
$$\text{標準誤差} = \text{平均値の標準偏差} = \sqrt{\frac{\text{不偏分散}}{\text{データ数}}} \quad (\text{もしくは、} = \frac{\text{不偏標準偏差}}{\sqrt{\text{データ数}}})$$

標準誤差のイメージ

まだイマイチぴんとこない標準誤差について、もう少し説明します。標準誤差は「平均値」の標準偏差、つまり平均値のバラツキを表します。ここでいう平均値とは、母集団を推定するために抽出した標本から計算した標本平均のことです。いま、牛群全体の体重が知りたい！ということにしましょう。500頭全部を調べることはしないので、30頭の標本を抽出し、牛群全体の平均体重を推定することにしました。



上の図は、30頭の標本抽出を4回繰り返した場合のイメージです。このように、平均値は当然サンプルした30頭のウシが全く同じでない限り、毎回違う値を示します。つまり、平均値がバラツクのです。しかし、**実際には1回の標本抽出**しかできませんよね。なので、平均値のバラツキを調べることが大事になってくるわけです。流れとしては、抽出をして平均値を求めますが、その1回のデータを使って不偏分散、さらには標準誤差（平均値の標準偏差：バラツキ）を計算するという感じです。



エクセルで求めるには、不偏分散は $VAR.S$ （選択したデータセル）、不偏標準偏差は $STDEV.S$ （選択したデータセル）となります。そして標準誤差は $STDEV.S$ （選択セル） / $SQRT(COUNT(選択セル))$ もしくは、 $SQRT(VAR.S(選択セル)/COUNT(選択セル))$ で計算することができます。

まとめ

標準誤差はうまくイメージがつかないと少し理解しづらいかもしれませんが、しかし、大事なのは母集団が知りたい！でも全頭は調べられないので標本を抽出して推定する（平均値などを計算する）！でも平均値は抽出した標本によって違うよね！だから平均のバラツキをチェックするために標準誤差も計算しよう！という流れです。しかし、これらはあくまで平均という1点の推定値です。そこで、次回は1点だけでなく、〇〇から△△までという区間についての推定に話を広げていきたいと思います。