ありがたや気力戻りてアイキャスト　＜不偏分散＞

2020 年 12 月中旬　白田由香利

IEEE iCAST 2020 の日が明日に迫った．自分がセッションをオーガナイズした，絶対に成功させたい特別セッションもあり私がこのセッションチェアをするのであるが，五十肩が痛くて気力が集中できない．前日になっても今ひとつ自分で乗り切れていない．どうしよう．それが当日の朝，今まで激痛で夜中目が覚めてしまっていたのだが，なんと朝まで眠れた．まだ腕が上がらないままで痛いのであるが，何か違う．毎日通っている整体の先生のところに朝一番で行ったが，先生も「からだがほぐれていますから今日はいけるのではないですか（先生も今日の国際会議のことを知っている）」．私もそう思った．プロたるもの，自分でオーガナイズしたセッションでは，発表者も参加者も新しいことを学び多いにディスカッションして満足して帰ってもらう，というミッションがある．国際会議の参加費はバーチャルといえども安くない．それには，まず自分が発表論文の内容を面白いと思い，意欲的に質問させて頂く，という情熱が必要である．整体から出て水分補給にカフェに入り，本日の論文を読み直す．頭がくるくる動いて，著者が主張したいことの輪郭がはっきりと見えてくる．質問もどしどし湧いてくる．質問したい，ご教示頂きたい，というパッションが体にみなぎってくる．これなら今日はいける，と感じた．果たして，皆様の活発なご議論があり，セッションは成功裏に終わることができた．崖っぷちに来て，からだが総動員して火事場の馬鹿力を発揮して肩の痛みを修復してくれたのではないかと思う．そして，利他の精神と気力体力回復，は鶏と卵の関係のような気がした．

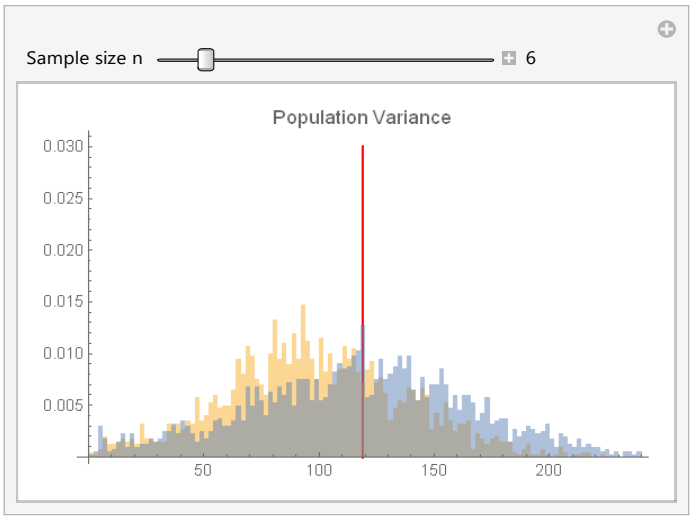
さて，不偏分散の公式の話をする．統計の授業での FAQ のひとつに，「どうしてデータサイズ n ではなくて，(n-1)で割るのですか？」というものがある．

答え：分散の定義は，**偏差の平方和を自由度**で割った量だからです．

学生：自由度とは何ですか？

答え：自由に値を動かすことができるデータ数です．

ビジュアルに説明をする．母集団から標本抽出を行う．母集団サイズが 10000 個あって，その中から標本サイズ 6 個の標本を抽出した．推定したい量は，後ろにある**母集団の平均（母平均）と母集団の分散(母分散)**である．2000 回のシミュレーション結果を以下に示す．母分散は赤の棒で示されている値である．実際のデータ分析では，この値が分からない．この値を推定するために，標本抽出を行い，その結果を使って推定しようとする．



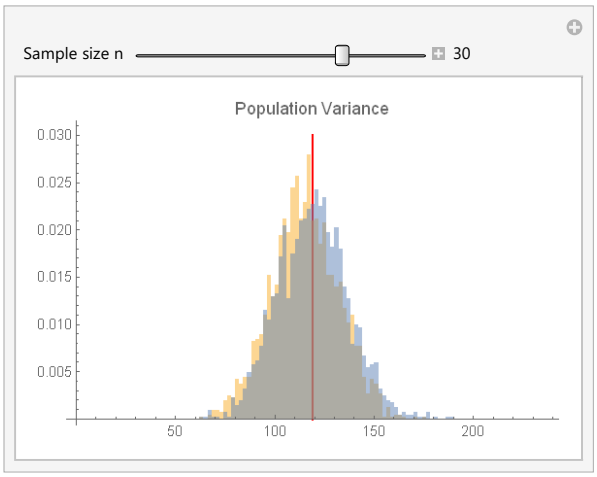
(1)　記述統計の分散　$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

(2)　不偏分散　$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$

$\bar{X}$ は標本平均である．$\bar{X} = \frac{x_1+x_2+\cdots+x_n}{n}$

標本サイズ n で割った場合と，(n－1)で割った場合の式を上記に示した．記述統計の分散のシミュレーションの結果（黄色）を見ると，真の母分散の値よりも小さくところに分布していることが見える．不偏分散のシミュレーション結果(青)

は，真の母分散値の周囲に大小に等分に分布している，つまり偏りがない，不偏に分布していることが見える．まずは，このシミュレーションから「母分散を推定しようとしたら，n-1 で割る不偏分散の式を使うべきだ」ということを納得して頂きたい．もう一つ，標本サイズ 30 の場合のシミュレーションの結果を示す(左図参照)．

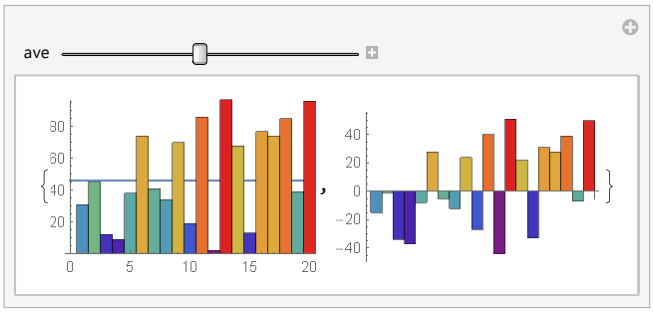


n=30 であると，30 と 29 の違いであるので，ヒストグラムの違いも小さくなってくる．しかし，よく見ると記述統計の分散のほうが，小さい値のほうに偏っていることが分かる．再度，ポイントを言う．

母分散を推定するときは，不偏分散の式を使う．

次に，自由度の話をする．どうして，標本において自由度が 1 減ったのか，ビジュアルに説明しよう．

標本サイズ 20 の標本をとってきた(下図左)．その標本平均$\bar{X}$を計算する．

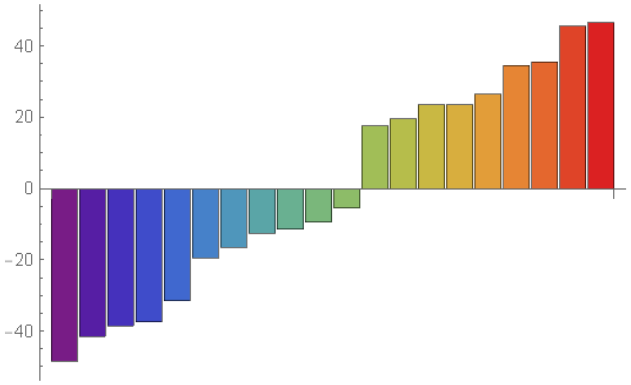


$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

母平均μの値が分からないので，μの代わりに$\bar{X}$を使って，$\sum_{i=1}^{n}(x_i - \bar{X})^2$　の項を計算する．

この標本の式が制約となってくる．

左図の右側は，偏差$x_i - \bar{X}$を示している．偏差をそのまま合計したら，0 になる．



$$\sum_{i=1}^{n}(x_i - \bar{X}) \quad = 0$$

偏差を小さい順にソートしたものが左の図である．マイナスの合計とプラスの合計が同じである．換言すると，標本平均からのずれ偏差の合計が 0 となるように，標本平均は計算されている．

ということは，1 番目のデータから n-1 番目のデータまでは自由に値をとれるが，最後のデータは，つじつま合わせをしなくてはならないので，自由に動くことができない．羊羹を分けるのに，最後の一人は余りもの分しかないのと似ている（しかし，前の人が取りすぎると，マイナスになってしまい，最後の人は自分で羊羹を補填しなくてはいけない)．

母平均を標本平均で代用した．標本平均は制約式となる．最後のデータは，自由に動けない．

よって，自由度は 1 減じた．

最後に，記述統計と推測総計の話をする．母集団を対象として，平均や分散などを計算する統計が記述統計である．母集団から標本を抽出して，標本の平均や分散を使って母集団の平均や分散などを推定するのが推測統計である．データ分析では，推測統計の場合が殆どである．よって Excel にしろ Mathematica にしろ，デフォルトの分散は，不偏分散である．Excel の場合，関数 var が不偏分散で，var.p が記述統計

の分散になる．Var.p の p は母集団 population の頭文字である．

終わり

引用元：ありがたや雪をかをらす南谷　芭蕉

参考：

IEEE iCAST 2020, Special session “Awareness Technology for Economic and Social Data Analysis,” Organizers: Prof. Yukari Shirota, Prof. Basabi Chakraborty, Prof. Takako Hashimoto, Qingdao, China and virtual, Dec. 7-9, 2020. http://www.icast2020.org.cn/ss/ss01.html

グラフィクス教材の場所：https://www-cc.gakushuin.ac.jp/~20010570/VDStat/