標本平均を用いた変動は必ず小さくなるか？

Sampo Suzuki, CC 4.0 BY-NC-SA

2021-06-27

標本平均を用いた変動は必ず小さくなるのか？

『ソフトウェアメトリクス統計分析入門』[小池, 2015] の 3.3 ワンポイント講義「不偏分散を算出する際に自由度を用いる理由」には、標本平均（$\bar{x}$）を使って算出した変動（$\sum_{i=1}^{n}(x_i - \bar{x})^2$、偏差平方和）は母平均（$\mu$）を使って算出した変動（$\sum_{i=1}^{n}(x_i - \mu)^2$）よりも必ず小さな値になるとあります。実際に小さくなるのかを確認します。

母集団データの作成

最初に正規分布を持つ母集団のデータ（x）を作成します。ここでは母平均と母標準偏差は不明であると仮定します。

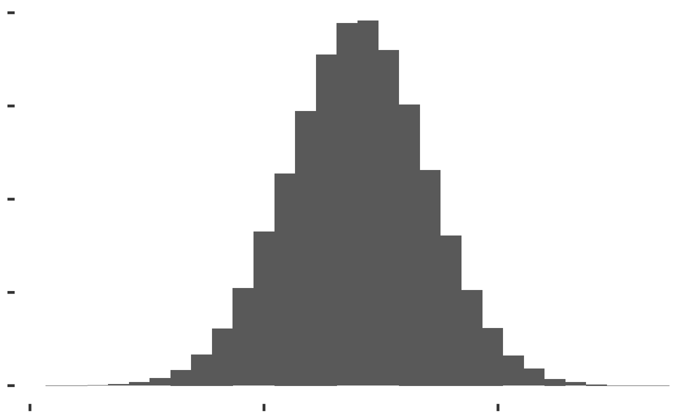

Figure 1: 母集団の分布

簡単なシミュレーション

上記の母集団（x）から以下の手順で二種類の変動（偏差平方和）を求めます。

1. 3 つのデータをランダムサンプリングで取り出す（標本 $x_n, n = 1, 2, 3$）
2. 取り出したデータの平均値（標本平均 $\bar{x}$）を求める
3. 標本平均（$\bar{x}$）を用いて標本の変動（偏差平方和 $\sum_{i=1}^{n}(x_i - \bar{x})^2$）を求める
4. 母平均（μ）を用いて標本の変動（偏差平方和 $\sum_{i=1}^{n}(x_i - \mu)^2$）を求める
5. 求めた二つの変動（偏差平方和）を比較する

この計算を任意の回数繰り返して標本平均（$\bar{x}$）を用いた標本の変動（偏差平方和 $\sum_{i=1}^{n}(x_i - \bar{x})^2$）の方が小さいことを確認します。

```r
df <- data.frame()
for (i in c(1:30)) {
  xs <- sample(x, size = 3, replace = FALSE)  # 母集団から 3 つのデータを取り出す
  xb <- mean(xs)                      # 標本平均を求める
  dssxb <- sum((xs - xb)^2)           # 標本平均を用いた変動（偏差平方和）
  dssmu <- sum((xs - mu)^2)           # 母平均を用いた変動（偏差平方和）
  # 計算結果をデータフレームにまとめる
  dftmp <- data.frame(no = i,         # 通し番号
                      x1 = xs[1],     # 標本データ（n = 1）
                      x2 = xs[2],     # 標本データ（n = 2）
                      x3 = xs[3],     # 標本データ（n = 3）
                      xb,             # 標本平均
                      mu,             # 母平均
                      dssxb,          # 標本平均を用いた変動（偏差平方和）
                      dssmu,          # 母平均を用いた変動（偏差平方和）
                      diff = dssxb - dssmu  # 負値なら標本平均による変動が小さい
                      )
  df <- dplyr::bind_rows(df, dftmp)
}

df %>%
  dplyr::rename(`標本平均`= xb, `母平均` = mu,
                `標本平均での変動` = dssxb, `母平均での変動` = dssmu,
                `変動差（標本-母）` = diff) %>%
  df_print(all = TRUE, scale_down = TRUE, caption = " シミュレーション結果")
```

Table 1: シミュレーション結果

no	x1	x2	x3	標本平均	母平均	標本平均での変動	母平均での変動	変動差（標本-母）
1	2.006349	4.9218712	13.662108	6.8634430	3.996745	73.5829117	98.236793	-24.6538812
2	1.788097	0.3882464	9.061179	3.7458409	3.996745	43.3590209	43.547879	-0.1888578
3	4.911329	1.6467477	4.132243	3.5634399	3.996745	5.8140515	6.377310	-0.5632586
4	7.012622	2.3197320	5.685805	5.0060530	3.996745	11.7047028	14.760814	-3.0561109
5	5.387746	3.5307648	5.427929	4.7821466	3.996745	2.3497420	4.200311	-1.8505694
6	2.101703	5.9993869	3.536314	3.8791345	3.996745	7.7722602	7.813756	-0.0414963
7	6.205521	4.1200703	4.464700	4.9300970	3.996745	2.4994440	5.112885	-2.6134408
8	1.739977	-3.7108190	9.146871	2.3920096	3.996745	83.2978089	91.023331	-7.7255225
9	4.071693	0.7495883	1.788127	2.2031361	3.996745	5.7765378	15.427630	-9.6510926
10	5.066624	4.3646527	5.665752	5.0323429	3.996745	0.8481924	4.065585	-3.2173922
11	3.162586	2.4739477	3.156060	2.9308644	3.996745	0.3131807	3.721481	-3.4083007
12	3.426702	4.7470220	7.605623	5.2597821	3.996745	9.1260758	13.911868	-4.7857920
13	4.858415	6.6613752	3.807375	5.1090549	3.996745	4.1668899	7.878594	-3.7117037
14	7.448547	7.1965942	0.596943	5.0806946	3.996745	30.1877836	33.712628	-3.5248440
15	5.204287	-4.9059256	1.280274	0.5262118	3.996745	51.9611147	88.094905	-36.1337908
16	2.670844	5.5704882	8.045905	5.4290791	3.996745	14.4756391	20.630386	-6.1547472
17	3.156763	-0.3528402	7.303819	3.3692471	3.996745	29.3799398	30.561199	-1.1812588
18	4.551711	1.5130258	3.577730	3.2141556	3.996745	4.8150824	6.652418	-1.8373358
19	8.398765	7.4489186	3.508472	6.4520516	3.996745	13.4480976	31.533696	-18.0855988
20	5.417183	5.9783885	5.468162	5.6212443	3.996745	0.1926273	8.109627	-7.9169993
21	2.735636	7.3445911	5.119664	5.0666305	3.996745	10.6254527	14.059421	-3.4339684
22	4.938459	4.7899228	1.256361	3.6615810	3.996745	8.6886552	9.025659	-0.3370036
23	1.475681	1.8195438	2.536436	1.9438869	3.996745	0.5857917	13.228464	-12.6426723
24	8.819532	0.2646054	6.147329	5.0771555	3.996745	38.3112911	41.813155	-3.5018640
25	4.115611	5.3069370	4.355087	4.5925452	3.996745	0.7942081	1.859144	-1.0649354
26	4.166215	4.7324917	2.513740	3.8041492	3.996745	2.6580668	2.769346	-0.1112788
27	2.464780	6.2689757	1.749776	3.4945107	3.996745	11.8020997	12.558816	-0.7567164
28	7.425591	2.2166859	4.767910	4.8033957	3.996745	13.5682362	15.520295	-1.9520586
29	3.258395	6.3141248	4.093102	4.5552074	3.996745	4.9890542	5.924697	-0.9356424
30	5.353118	1.2392623	4.160559	3.5843131	3.996745	8.9599938	9.470293	-0.5102988

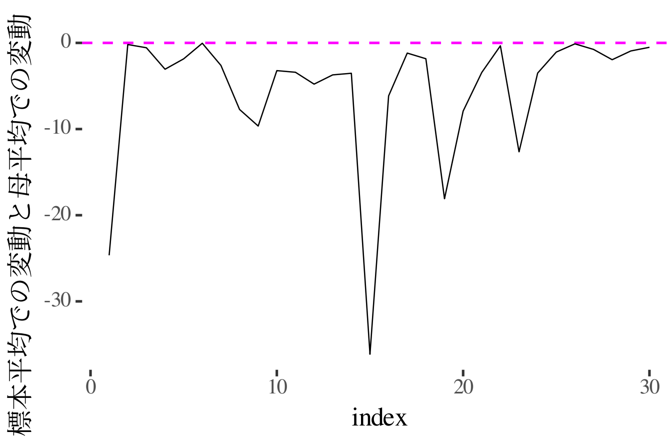

Figure 2: 標本平均を用いた変動と母平均を用いた変動の差

標本標準偏差の補正値を確認する

標本平均（$\bar{x}$）を用いた変動（偏差平方和）は母平均（μ）を用いた変動（偏差平方和）よりも小さくなることがわかりました。では、自由度で補正した標準偏差が母標準偏差に本当に近くなるのかを同じ母集団（x）を使って確認します。

```
df <- data.frame()
m <- 12
for (i in c(1:30)) {
  xs <- sample(x, size = m, replace = FALSE)  # 母集団から 3 つのデータを取り出す
  xb <- mean(xs)                              # 標本平均を求める
  sdxb <- sqrt(sum((xs - xb)^2) / (m - 1))    # 自由度で補正した標本標準偏差
  sdmu <- sqrt(sum((xs - mu)^2) / m)          # 母標準偏差
  # 計算結果をデータフレームにまとめる
  dftmp <- data.frame(no = i,             # 通し番号
                      x1 = xs[1],         # 標本データ (n = 1)
                      x2 = xs[2],         # 標本データ (n = 2)
                      x3 = xs[3],         # 標本データ (n = 3)
                      xb,                 # 標本平均
                      mu,                 # 母平均
                      sdxb,               # 補正した標本標準偏差（不偏推定値）
                      sdmu,               # 母標準偏差
                      diff = sdxb - sdmu  # 負値なら標本平均による変動が小さい
                      )
  df <- dplyr::bind_rows(df, dftmp)
}

df %>%
  dplyr::rename(`標本平均`= xb, `母平均` = mu,
                `補正した標本標準偏差` = sdxb, `母標準偏差` = sdmu,
                `差（標本-母）` = diff) %>%
  df_print(all = TRUE, scale_down = TRUE, caption = "シミュレーション結果")
```

Table 2: シミュレーション結果

no	x1	x2	x3	標本平均	母平均	補正した標本標準偏差	母標準偏差	差（標本-母）
1	5.9005759	7.1291092	3.8242538	3.614148	3.996745	2.586970	2.506211	0.0807592
2	6.0248282	5.8620663	2.2198528	4.603901	3.996745	2.607905	2.569638	0.0382661
3	6.4648030	3.5847104	2.5342685	5.089606	3.996745	2.494076	2.626098	-0.1320228
4	-0.8838906	4.3998278	1.7926698	4.800031	3.996745	3.038578	3.018081	0.0204970
5	-4.5942678	1.0893745	0.4178212	2.565047	3.996745	3.663974	3.788897	-0.1249226
6	5.3967884	3.8588633	2.1832539	4.261000	3.996745	2.706099	2.604334	0.1017651
7	6.8730200	7.2502373	5.6432586	5.456434	3.996745	3.830144	3.946922	-0.1167784
8	0.5408073	4.6170769	5.8160347	3.987181	3.996745	2.642547	2.530064	0.1124828
9	-0.1232023	1.7702919	2.8303924	2.410388	3.996745	2.112865	2.570741	-0.4578755
10	3.0938152	3.4629398	6.1279842	3.502391	3.996745	2.252870	2.212884	0.0399858
11	4.3994310	7.0916869	3.1375773	3.809455	3.996745	2.191099	2.106162	0.0849376
12	6.2870186	9.1389105	0.0862662	4.409378	3.996745	2.810322	2.722135	0.0881873
13	6.8602403	10.4291384	9.6602453	4.463471	3.996745	3.401384	3.289852	0.1115316
14	9.6864373	2.1238370	1.2698589	4.606710	3.996745	3.791487	3.680963	0.1105247
15	6.3649314	0.9749754	3.4152821	4.939895	3.996745	2.568665	2.633958	-0.0652933
16	0.3682559	2.8779801	5.5504674	4.240597	3.996745	4.441971	4.259849	0.1821222
17	5.8434016	5.2180801	1.6120696	4.306186	3.996745	2.870485	2.765646	0.1048390
18	4.5487401	0.9673316	6.2789387	2.854302	3.996745	2.001293	2.230826	-0.2295327
19	3.6510340	5.0247488	2.7621445	3.788562	3.996745	1.957817	1.885993	0.0718247
20	5.6092354	7.7755369	2.8950185	5.226601	3.996745	2.121617	2.374595	-0.2529779
21	6.2705511	8.0279919	0.4769332	4.696590	3.996745	3.965896	3.861013	0.1048832
22	-3.1034010	2.6415654	5.6015774	1.792153	3.996745	3.843142	4.289423	-0.4462814
23	1.9884340	1.4358724	4.4602812	3.354269	3.996745	1.940232	1.965596	-0.0253637
24	2.2377706	2.9119022	-0.9604801	2.612762	3.996745	2.432963	2.709509	-0.2765463
25	10.1334809	9.3962861	0.9661430	4.402635	3.996745	3.630622	3.499673	0.1309489
26	1.4913636	2.2132959	3.8690116	3.775323	3.996745	2.509745	2.413078	0.0966669
27	0.6061874	5.7470604	6.8969015	3.734887	3.996745	2.320765	2.237340	0.0834250
28	0.7757434	4.4510336	6.5881741	4.846286	3.996745	2.941545	2.941657	-0.0001128
29	4.1562629	-0.1827302	4.1072400	3.105353	3.996745	2.683109	2.719142	-0.0360325
30	5.7547576	4.5193953	6.3142889	3.847106	3.996745	3.060298	2.933831	0.1264671

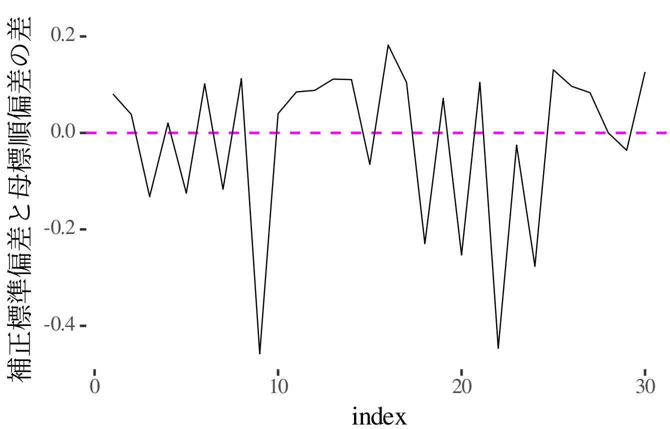

Figure 3: 標本標準偏差の補正値（不偏推定値）と母標準偏差の差

おわりに

詳細で理論的な説明が必要な場合は『なぜ不偏分散は N-1 で割るのか』[est] を参照してください。

ちなみに母集団（x）の平均値（mu）と標準偏差（s）は以下の通りでした。

```
mean(x)                  # 平均値
```

```
## [1] 3.996744
```

```
(n / (n - 1)) * sd(x)   # 標準偏差
```

```
## [1] 2.992983
```

References

なぜ不偏分散は n-1 で割るのか. URL http://kosugitti.sakura.ne.jp/wp/wp-content/uploads/2013/08/est.pdf.

利和小池. 『ソフトウェアメトリクス統計分析入門』. 日科技連出版, first edition, 2015. URL https://www.juse-p.co.jp/products/view/545. ISBN 978-4-8171-9558-6.