Re-Weighting Stratified Sampling Example

Marcelo Avila

29 November 2018

```r
library(magrittr)

## generate fake data
set.seed(42)
shape = 5    # shape = alpha param
location = 1000 # location = lower bound

popAll <- EnvStats::rpareto(10000,  location = location, shape = shape)

## divide into thirds
b1 <- 1080 # found adhoc so that divides into thirds
b2 <- 1250

# plot
hist(popAll, col=rgb(1,0,0,0.5), freq = T, breaks = 100, xlim = c(1000, 3000))
abline(v = b1, lty=3, lwd=2, col = "blue")
abline(v = b2, lty=3, lwd=2, col = "blue")

mean(popAll<b1)
```

```
## [1] 0.3223
```

```r
mean(popAll>b2)
```

```
## [1] 0.3288
```

```r
mean(popAll>b1 & popAll<b2)
```

```
## [1] 0.3489
```

```r
# divide into 3 strata
p1 <- popAll[popAll<=b1]
p2 <- popAll[popAll>b1 & popAll<=b2]
p3 <- popAll[popAll>b2]

# check means
mean1 <- mean(p1);mean1
```

```
## [1] 1036.133
```

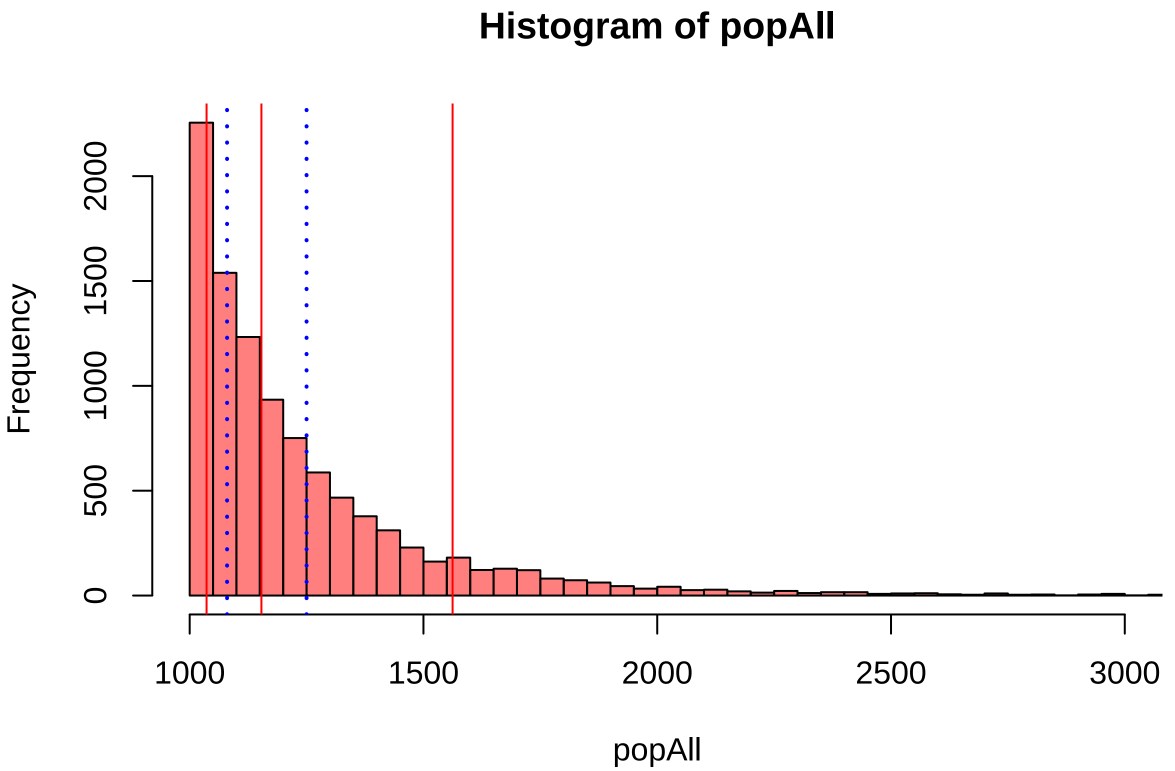
```r
abline(v=mean1, col="red")
mean2 <- mean(p2);mean2
```

```
## [1] 1153.442
```

```r
abline(v=mean2, col="red")
mean3 <- mean(p3);mean3
```

```
## [1] 1562.301
```

```r
abline(v=mean3, col="red")
```




```r
# check consistency
sum(p1, p2, p3) == sum(popAll)
```

```
## [1] TRUE
```

```r
max(p1) < min(p2); max(p2) < min(p3)
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```r
# ###########################
# STRATIFICATION METHOD
# ###########################

# Disproportional Strata
gew1 <- 1/7; gew2 <- 2 * gew1; gew3 <- 4 * gew1


ss <- 700      # total sample size
s1 <- sample(p1, size = gew1 * ss);s1;length(s1) # strata 1
```

```
##    [1] 1069.024 1059.288 1069.182 1028.192 1029.749 1056.177 1033.158
##    [8] 1027.798 1038.760 1017.764 1020.261 1002.459 1047.531 1075.483
##   [15] 1005.797 1039.770 1068.835 1079.889 1070.217 1025.327 1024.438
##   [22] 1030.499 1018.250 1047.707 1017.243 1014.171 1061.029 1005.201
##   [29] 1002.859 1075.515 1075.329 1073.046 1073.489 1030.534 1023.797
##   [36] 1003.377 1052.865 1020.320 1018.821 1067.738 1011.351 1044.467
##   [43] 1017.677 1010.539 1001.522 1064.941 1018.319 1010.356 1021.136
##   [50] 1052.266 1007.049 1002.888 1075.266 1063.035 1031.304 1009.697
```

```
##  [57] 1028.638 1026.437 1018.717 1070.147 1011.264 1046.094 1004.161
##  [64] 1022.703 1002.216 1001.868 1050.022 1018.111 1025.896 1052.609
##  [71] 1060.672 1065.286 1013.405 1006.320 1041.417 1001.606 1021.364
##  [78] 1040.138 1068.367 1001.412 1040.441 1035.202 1053.509 1045.181
##  [85] 1073.411 1030.660 1047.370 1066.075 1079.896 1004.357 1000.694
##  [92] 1062.475 1027.274 1050.770 1059.178 1033.183 1004.640 1010.647
##  [99] 1006.767 1076.351
```

```
## [1] 100
```

```r
s2 <- sample(p2, size = gew2 * ss);s2;length(s2) # strata 2
```

```
##   [1] 1220.815 1086.533 1181.470 1248.771 1246.300 1155.622 1165.273
##   [8] 1122.024 1233.013 1118.006 1085.347 1096.227 1114.080 1131.932
##  [15] 1146.753 1193.157 1103.662 1147.301 1175.141 1212.075 1152.179
##  [22] 1090.442 1178.715 1176.248 1221.335 1156.726 1132.400 1090.872
##  [29] 1097.886 1139.427 1137.543 1143.338 1206.445 1149.890 1123.965
##  [36] 1157.360 1235.342 1124.969 1157.920 1227.593 1244.880 1108.290
##  [43] 1097.362 1145.689 1086.648 1080.184 1101.022 1163.064 1152.904
##  [50] 1163.113 1098.313 1120.393 1229.213 1108.833 1111.166 1102.129
##  [57] 1153.172 1126.451 1165.627 1101.202 1203.828 1082.814 1124.797
##  [64] 1167.521 1081.720 1146.361 1122.194 1131.310 1121.250 1151.382
##  [71] 1122.352 1232.088 1141.745 1168.925 1192.940 1088.551 1104.623
##  [78] 1093.289 1080.132 1201.362 1088.165 1117.526 1118.672 1102.694
##  [85] 1089.246 1163.878 1124.310 1097.451 1247.084 1091.075 1172.339
##  [92] 1131.130 1156.773 1132.751 1194.125 1249.072 1146.757 1142.938
##  [99] 1187.387 1125.514 1234.450 1169.500 1119.773 1111.032 1157.426
## [106] 1180.522 1194.750 1097.643 1111.140 1242.216 1103.529 1165.388
## [113] 1201.227 1200.966 1180.567 1147.188 1090.920 1129.867 1230.044
## [120] 1084.529 1219.661 1098.471 1107.244 1131.037 1129.500 1244.713
## [127] 1191.445 1165.096 1084.387 1135.664 1226.971 1084.136 1153.397
## [134] 1212.910 1080.515 1190.800 1203.684 1099.751 1234.818 1082.720
## [141] 1107.298 1150.871 1117.039 1202.578 1206.285 1088.590 1161.726
## [148] 1239.926 1196.513 1174.380 1140.550 1171.823 1132.922 1228.079
## [155] 1080.850 1142.195 1169.560 1096.082 1223.693 1129.167 1184.557
## [162] 1107.914 1085.361 1125.476 1244.175 1181.242 1115.564 1108.489
## [169] 1181.668 1209.016 1163.226 1094.766 1109.945 1093.174 1166.387
## [176] 1140.049 1150.400 1186.103 1186.453 1186.559 1105.499 1207.702
## [183] 1112.308 1096.913 1088.583 1159.915 1170.794 1110.848 1161.663
## [190] 1225.720 1086.606 1135.319 1085.227 1080.105 1211.156 1218.242
## [197] 1212.503 1145.231 1135.900 1143.750
```

```
## [1] 200
```

```r
s3 <- sample(p3, size = gew3 * ss);s3;length(s3) # strata 3
```

```
##   [1] 2182.720 2000.903 1452.030 1359.156 1348.681 1284.319 1643.376
##   [8] 1266.442 1252.651 1642.368 1473.784 1440.134 1682.191 1662.439
##  [15] 1439.643 1331.449 1312.069 1277.481 1346.570 1308.026 1367.818
##  [22] 1943.113 1759.794 2053.953 2402.872 1284.657 1707.024 1311.636
##  [29] 1303.038 1299.252 1410.337 1738.883 1574.058 1513.286 1436.774
##  [36] 1369.393 1355.320 1307.952 1398.737 1296.271 1371.617 1673.456
##  [43] 1553.757 1731.687 1362.450 1476.493 1401.093 1408.079 1416.764
##  [50] 1464.004 1557.801 1307.892 3269.475 1548.311 1344.627 1295.782
##  [57] 1545.850 1777.429 1867.790 1719.611 2131.888 1665.194 1445.159
##  [64] 1981.197 1489.221 1314.013 1325.177 1664.860 1323.720 1596.449
```

```
##  [71] 1916.707 2121.884 1265.109 1741.123 5214.613 1257.908 1683.444
##  [78] 1729.266 1367.213 1354.445 1922.972 1459.490 1342.073 1367.736
##  [85] 1379.208 1773.529 1679.178 1465.960 1463.266 1258.785 1271.978
##  [92] 1295.700 1499.434 1325.674 1262.633 1711.144 1340.688 1602.938
##  [99] 1273.775 1486.074 1282.896 1400.895 1448.082 1446.260 4242.931
## [106] 1322.104 1476.393 1612.099 1258.145 1425.183 1408.029 1401.692
## [113] 1254.023 1413.936 1478.654 1368.189 1570.666 1613.134 1259.003
## [120] 1640.433 2066.469 1604.176 2254.142 1754.209 1416.978 1372.460
## [127] 1368.590 1445.554 1393.912 1339.994 2014.330 1383.148 1253.055
## [134] 1392.151 4324.845 1304.235 1908.268 1254.933 1846.784 1330.792
## [141] 1377.397 1466.908 1363.532 1768.660 1302.079 1283.006 1347.554
## [148] 1279.976 1358.148 1911.630 2504.087 1309.012 1873.674 1701.689
## [155] 1477.236 1270.699 1748.282 1914.462 1297.779 1306.214 1385.131
## [162] 1272.191 1271.785 1352.658 1269.331 1261.867 1834.891 1364.916
## [169] 1344.828 1689.475 1313.028 1434.954 2253.612 1488.377 1451.860
## [176] 1250.924 1800.937 1270.793 1402.829 1987.239 1478.370 1383.787
## [183] 1473.812 1647.103 1328.201 1417.325 1425.275 1425.771 1270.967
## [190] 1418.334 1466.994 1496.551 2471.790 1364.202 1732.802 1264.238
## [197] 2208.731 1561.961 2470.036 1713.266 2459.635 2054.451 1706.850
## [204] 1595.508 1257.351 1470.120 1433.902 1499.508 1309.086 1267.462
## [211] 1411.504 1348.767 1807.593 1288.195 1268.461 1347.423 1436.876
## [218] 2258.415 2570.899 1444.287 1304.477 1287.688 1831.013 1693.744
## [225] 1650.445 1722.916 1470.342 1269.794 1357.523 1370.958 1871.956
## [232] 1304.602 2258.399 1427.152 1940.082 2116.936 1420.809 1400.007
## [239] 1422.683 1721.569 2508.204 1299.330 1424.961 1379.734 1631.710
## [246] 1256.376 1262.130 1477.423 1437.135 1660.301 1620.720 1342.513
## [253] 1260.152 1253.816 1294.696 1741.393 1359.528 1308.546 1251.787
## [260] 1282.131 1548.842 1802.206 1351.462 1258.815 1373.523 1839.521
## [267] 2132.475 1350.598 1295.945 1371.858 1342.258 1275.449 1389.246
## [274] 2133.194 1672.140 1760.364 1337.432 1282.115 1256.799 1858.533
## [281] 1252.278 1721.693 1417.348 1306.208 1868.816 1581.892 1436.649
## [288] 1358.054 1805.611 1302.776 2310.550 1252.523 1293.724 1328.231
## [295] 1438.324 2553.893 1404.898 1527.123 1824.719 1364.324 1597.137
## [302] 1314.058 1277.587 1564.263 1362.834 1289.120 1427.009 1593.583
## [309] 1514.134 1733.443 1546.452 1376.577 3053.292 1273.348 1369.649
## [316] 1504.447 1842.321 1350.195 1367.456 1518.552 1373.771 1790.336
## [323] 1360.201 1298.807 1284.071 1286.726 2063.301 1276.600 1266.537
## [330] 1365.025 1367.996 1605.368 1471.526 1327.886 1480.519 1269.612
## [337] 1273.538 1278.544 3091.121 1278.516 1335.633 1303.330 1340.571
## [344] 1365.778 1389.660 1252.344 1872.002 1272.889 1534.568 2743.428
## [351] 1979.366 1298.209 1487.240 1740.526 1765.163 1601.423 1396.953
## [358] 1831.802 1273.089 2073.492 1733.124 1770.107 1927.540 1806.449
## [365] 1479.230 1334.075 1655.707 1273.166 1322.372 1251.110 1315.129
## [372] 1387.825 1373.001 1737.953 1272.162 1250.340 1378.145 1295.729
## [379] 1570.986 5682.612 1380.712 1421.416 1498.311 1830.741 1286.944
## [386] 1320.238 1288.503 1330.633 1556.497 1256.964 1416.653 1398.593
## [393] 1266.573 1390.315 1546.718 1668.912 1330.319 2397.548 1755.459
## [400] 1445.937
```

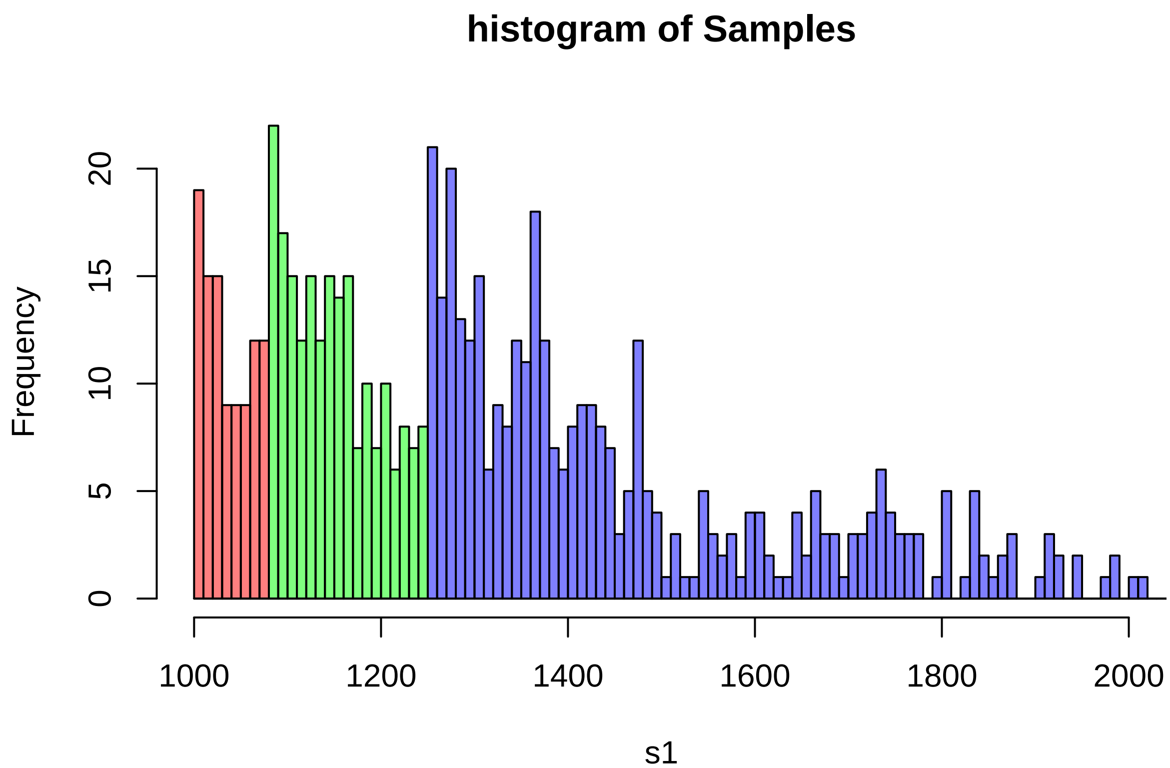
```
## [1] 400
```

```r
#check consistency (overlapping)
max(s1) < min(s2); max(s2) < min(s3)
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
breaks = 100:10000 * 10
hist(s1, col=rgb(1,0,0,0.5),
     xlim = c(1000, 2000), ylim = c(0,22), breaks = breaks,
     main = "histogram of Samples")
hist(s2, col=rgb(0,1,0,0.5), breaks = breaks,
     add=T);
hist(s3, col=rgb(0,0,1,0.5), breaks = breaks,
     add=T);
```




```
mean(s1);mean(s2);mean(s3)
```

```
## [1] 1035.497
## [1] 1149.43
## [1] 1566.506
```

```
mean(c(p1,p2,p3))    # mean of pops
```

```
## [1] 1250.066
```

```
mean(c(s1,s2,s3))    # mean of samps (wrong!)
```

```
## [1] 1371.483
```

```
sum(mean(s1), mean(s2), mean(s3)) / 3  # (sum of means, also wrong!)
```

```
## [1] 1250.478
```

```
length(p1); length(s1)
```

```
## [1] 3223
```

```
## [1] 100
```

```r
length(p2); length(s2)
```

```
## [1] 3489
## [1] 200
```

```r
length(p3); length(s3)
```

```
## [1] 3288
## [1] 400
```

```r
prop1 <- length(p1) / length(s1); prop1
```

```
## [1] 32.23
```

```r
prop2 <- length(p2) / length(s2); prop2
```

```
## [1] 17.445
```

```r
prop3 <- length(p3) / length(s3); prop3
```

```
## [1] 8.22
```

```r
# sample and pop sizes
N1 <- length(p1);
N2 <- length(p2);
N3 <- length(p3)
N <- sum(N1, N2, N3);N
```

```
## [1] 10000
```

```r
sum(c(mean(s1) * ((N1 * prop1) / N),
      mean(s2) * ((N2 * prop2) / N),
      mean(s3) * ((N3 * prop3) / N))
    )
```

```
## [1] 21986.39
```

```r
# unbiased estimation
hatMEAN <- (1/N) * sum( c(N1,N2,N3 ) * c( mean(s1), mean(s2), mean(s3) ))
reaMEAN <- mean(popAll)
hatMEAN;reaMEAN
```

```
## [1] 1249.844
## [1] 1250.066
```

```r
# OR
(N1 * mean(s1) + N2 * mean(s2) + N3 * mean(s3)) / N
```

```
## [1] 1249.844
```

```r
wgt1 <- N1 / N;    wgt2 <- N2 / N;    wgt3 <- N3 / N

barY <- wgt1 * mean(s1) + wgt2 * mean(s2) + wgt3 * mean(s3); barY
```

```
## [1] 1249.844
```

```r
# Interesting to see: its not the inverse of gew (1/7, 2/7 and 4/7) that gets
# to rewheight the averages of the strata mean into the overall mean, but the
# proportion of of observations in each strata compared to the whole poplation
```

```r
# (n1 / N) and they are all around 1/3 and not related to the specific wheights


### Standard error ((don't take it seriously))

sd(popAll)
```

```
## [1] 320.6379
```

```r
Nh <- sum(N1, N2, N3) # sum of all strata sizes
n1 <- length(s1);  n2 <- length(s2); n3 <- length(s3)
N1 <- length(p1);  N2 <- length(p2); N3 <- length(p3)
w1 <- N1 / N;      w2 <- N2 / N;     w3 <- N3 / N

Nv <- c(N1, N2, N3)
nv <- c(n1, n2, n3)
varS <- unlist(lapply(list(s1,s2,s3), var))

## within strata variance (needs checking, dont take seriously)
sum(Nv^2 * ((Nv-nv)/Nv) * varS/nv) / (Nh^2)
```

```
## [1] 52.48965
```

```r
var(popAll)
```

```
## [1] 102808.7
```

```r
# within variance much smaller, (maybe too muuch)
```