CS544, Fundamentals of Analysis Homework 5

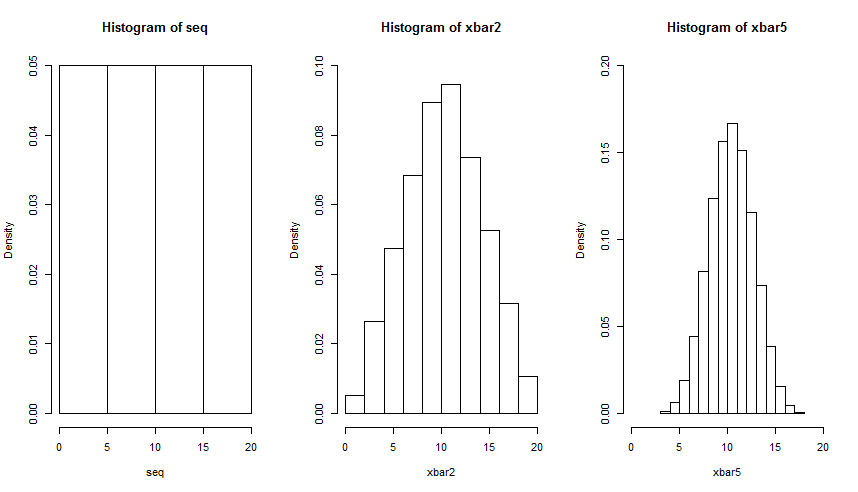
Jefferson Parker

[japarker@bu.edu](mailto:japarker@bu.edu)

**Part 1, Central Limit Theorem**

For the sequence of numbers 1 to 20, show the three plots in a single row:

1. Histogram of the densities of the distribution
2. All samples of size 2, histogram of densities of sample means
3. All samples of size 5, histogram of densities of sample means



1. Compare the means and standard deviations of the three distributions.

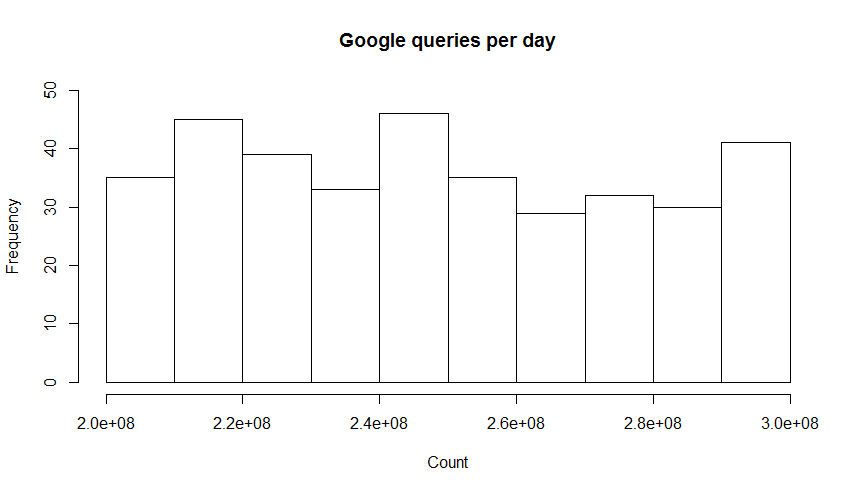
All three distributions have the same mean of 10.5, but the standard deviations are inversely proportional to the sample size. For the first distribution (density for n = 1)  = 5.92, the second (n = 2)  = 3.98 and for the third (n = 5)  = 2.29. This shows the increasing precision of the estimate of the mean with increasing sample size.

**Part 2, Central Limit Theorem**

Using the Google query counts for one year data in the queries.csv file:

1. Show the histogram for the distribution of the number of queries, the mean and the standard deviation.

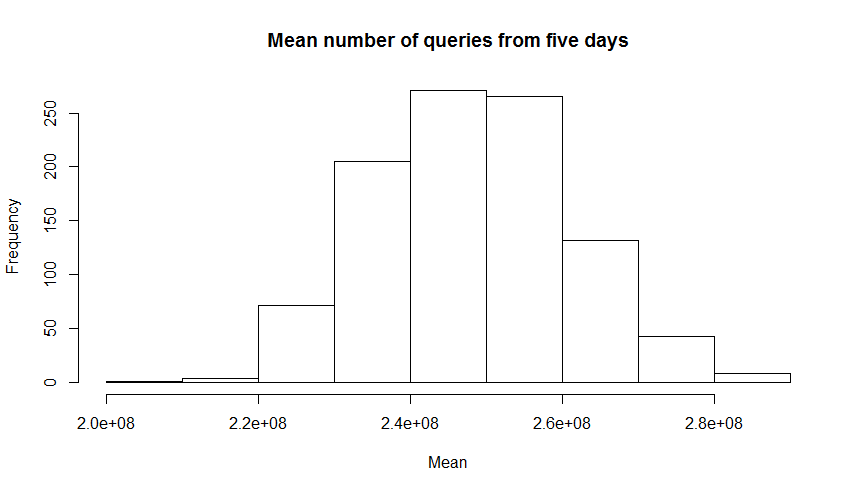
The histogram of the query data per day is:



The data has mean = 248,514,980 and standard deviation  = 29,202,674.

1. Draw 1000 samples of size n = 5, show the histogram of the densities of the sample mean and compute the mean and standard deviation of the sample means.

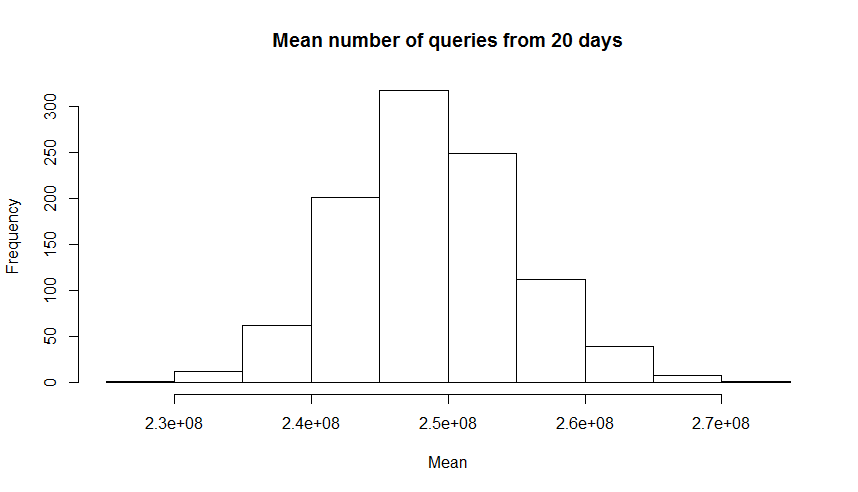
The histogram of the sample means is:



The n= 5 data has mean = 248,411,295 and standard deviation  = 12,873,958.

1. Draw 1000 samples of size n = 20, show the histogram of the densities of the sample mean and compute the mean and standard deviation of the sample means.

The histogram of the sample means is:



The n= 20 data has mean = 248,770,096 and standard deviation  = 6,318,110.

1. Compare the means and standard deviations of the three distributions.

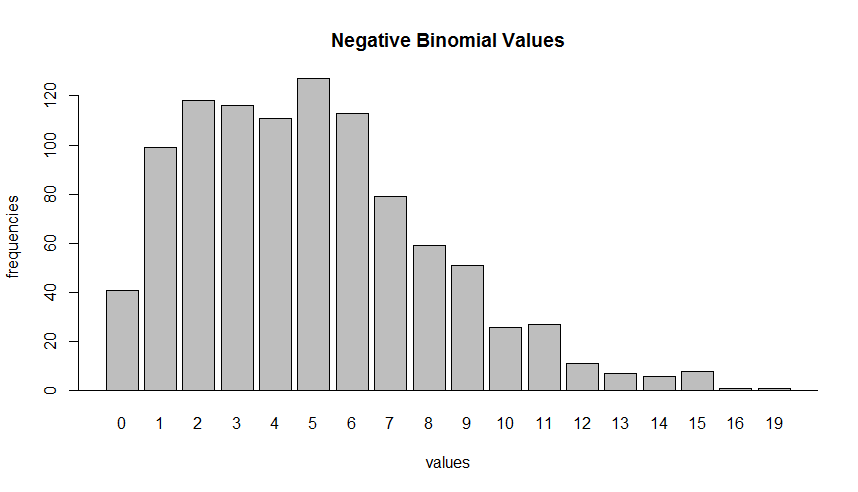
All three distributions have similar means, all being on the order of 248 million queries per day. The standard deviations show an inversely proportional relationship to the sample size, shrinking by almost half per distribution as the sample size increases to 20.

**Part 3, Central Limit Theorem – Negative Binomial Distribution**

For the input data following a negative binomial distribution with size = 5 and probability = 0.5:

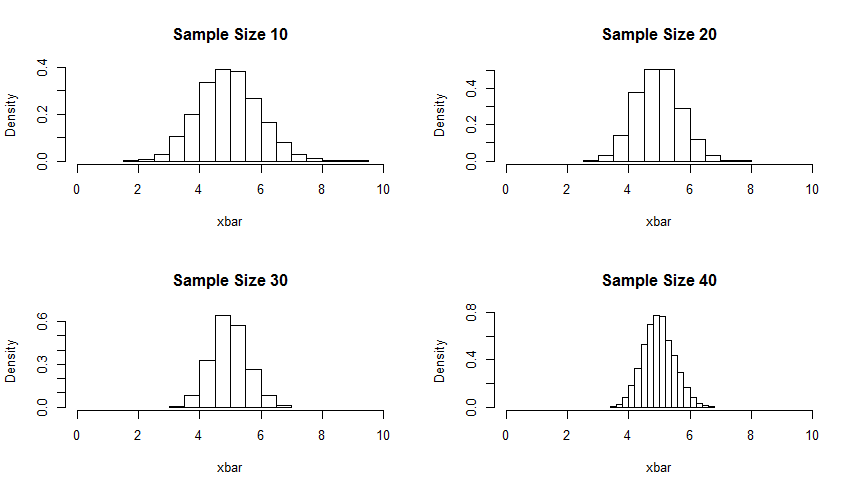
1. Generate 1000 random numbers; show the bar plot with the proportions of distinct values.

The barplot for 1000 negative binomial values is:



1. For samples of n = 10, 20, 30 and 40, generate data for 5000 samples from the same negative binomial distribution, show the histograms of the densities of their sample means in a 2 x 2 layout.

The histograms of densities of the four sample means are:



1. Compare the means and standard deviations of the data from a) with the four samplings from b).

The means and standard deviations of the distributions are:

|  |  |  |
| --- | --- | --- |
| Sample | Mean | Standard Deviation |
| 1000 values, n = 5 | 5.048 | 3.09 |
| 5000 values, n = 10 | 4.989 | 0.98 |
| 5000 values, n = 20 | 4.987 | 0.71 |
| 5000 values, n = 30 | 4.99 | 0.59 |
| 5000 values, n = 40 | 4.996 | 0.51 |

As is clear from the data, the mean value is about 5 for all distributions. Also, as the sample size increases the standard deviation shrinks.

**Part 4, Sampling**

Using the MU284 data set with sample size = 20:

1. Using simple random sampling without replacement, the frequencies of the Region variable (REG) and percentages of the total are shown in the table below.

The sample drawn was:

LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL

21 21 38 36 339 21 11 51 2055 4438 1 5

27 27 9 10 74 3 20 31 490 1751 2 6

52 52 7 6 38 6 11 35 280 767 3 10

61 61 18 18 129 8 17 49 970 1652 3 11

62 62 18 19 119 9 19 41 963 1470 3 11

92 92 12 11 73 5 28 41 492 1222 4 16

121 121 34 35 240 12 30 51 2089 3475 4 21

122 122 11 11 67 6 20 49 469 1726 5 22

133 133 13 10 82 9 15 41 576 2092 5 24

137 137 424 446 6720 21 35 81 47074 38945 5 24

140 140 15 15 118 7 23 41 781 3397 5 25

143 143 5 5 35 6 12 41 253 538 5 26

150 150 11 10 62 8 14 41 414 1119 5 27

153 153 11 10 68 10 14 41 474 1224 5 28

156 156 49 50 396 11 34 61 2881 4798 5 28

160 160 13 13 95 7 20 41 731 1101 5 28

176 176 13 13 83 5 19 41 623 1192 5 31

198 198 7 7 50 1 18 35 325 688 2 35

234 234 12 12 70 3 18 41 579 1150 6 42

264 264 4 4 28 2 19 31 199 422 8 47

It is worth noting that Region was not sampled at all in this sample draw.

The frequency of each region in the sample, and the percentage of the region in the sample and total data set are shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Frequency | % Sample | % Total |
| 1 | 1 | 5 | 8.8 |
| 2 | 2 | 10 | 16.9 |
| 3 | 3 | 15 | 11.3 |
| 4 | 2 | 10 | 13.4 |
| 5 | 10 | 5 | 19.7 |
| 6 | 1 | 5 | 14.4 |
| 7 | 0 | 0 | 5.2 |
| 8 | 1 | 5 | 10.2 |

1. Using systematic sampling, the sample drawn is:

LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL

5 5 56 52 536 20 27 61 3951 5183 1 1

19 19 27 28 250 9 22 41 1616 2208 1 4

33 33 65 62 488 14 33 61 3254 6389 2 7

47 47 118 119 1008 22 45 85 7619 12112 2 9

61 61 18 18 129 8 17 49 970 1652 3 11

75 75 13 13 90 8 24 49 538 2042 3 13

89 89 15 15 97 6 24 41 735 1321 4 16

103 103 17 15 122 14 17 41 894 1670 4 18

117 117 105 102 815 20 31 65 6323 9371 4 21

131 131 17 14 129 10 16 41 940 3690 5 23

145 145 23 22 166 9 26 49 1122 2025 5 26

159 159 22 21 146 12 15 49 1053 2066 5 28

173 173 18 17 128 11 19 49 1058 1592 5 31

187 187 13 14 86 7 15 41 552 1444 6 33

201 201 12 11 87 7 23 45 552 1408 2 36

215 215 14 15 112 7 22 41 761 1238 2 38

229 229 25 27 197 5 28 49 1239 2461 6 41

243 243 28 27 231 8 22 49 1560 2414 7 44

257 257 3 4 21 5 11 31 173 347 8 46

271 271 8 8 55 3 25 41 462 690 8 49

As expected, all regions are represented in this sample.

The frequency of each region in the sample, and the percentage of the region in the total data set are shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Frequency | % Sample | Percentage |
| 1 | 2 | 10 | 8.8 |
| 2 | 4 | 20 | 16.9 |
| 3 | 2 | 10 | 11.3 |
| 4 | 3 | 15 | 13.4 |
| 5 | 4 | 20 | 19.7 |
| 6 | 2 | 10 | 14.4 |
| 7 | 1 | 5 | 5.2 |
| 8 | 2 | 10 | 10.2 |

1. Using systematic sampling, but with inclusion probabilities based on the S82 variable, the sample drawn is:

LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL

13 13 53 40 386 24 13 51 2780 5931 1 3

25 25 21 19 164 8 25 45 1217 2389 1 5

39 39 6 6 37 8 17 41 257 655 2 8

52 52 7 6 38 6 11 35 280 767 3 10

66 66 20 19 127 9 20 49 956 1983 3 12

79 79 28 28 200 10 31 59 1499 5717 3 14

92 92 12 11 73 5 28 41 492 1222 4 16

107 107 21 19 145 12 25 49 1100 4117 4 19

120 120 24 24 172 13 23 49 1299 2355 4 21

135 135 11 10 65 6 17 41 418 1135 5 24

149 149 12 12 90 6 19 45 534 1243 5 27

163 163 7 5 39 9 13 35 329 637 5 29

179 179 9 10 63 6 22 41 411 869 6 32

194 194 8 9 63 3 23 41 411 950 2 35

209 209 17 19 132 7 31 49 989 1197 2 37

224 224 20 18 153 4 19 49 1039 2223 6 40

238 238 31 32 233 5 32 51 1579 3298 6 43

250 250 14 13 94 5 23 49 813 2486 7 45

267 267 4 5 35 2 19 31 252 687 8 48

282 282 29 27 226 7 28 49 1682 2898 8 50

The frequency of each region in the sample, and the percentage of the region in the total data set are shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Frequency | % Sample | Percentage |
| 1 | 2 | 10 | 8.8 |
| 2 | 3 | 15 | 16.9 |
| 3 | 3 | 15 | 11.3 |
| 4 | 3 | 15 | 13.4 |
| 5 | 3 | 15 | 19.7 |
| 6 | 3 | 15 | 14.4 |

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Frequency | % Sample | Percentage |
| 7 | 1 | 5 | 5.2 |
| 8 | 2 | 10 | 10.2 |

1. Using stratified sampling on the ordered REG variable, the sample drawn is:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LABEL | P85 | P75 | RMT85 | CS82 | SS82 | S82 | ME84 | REV84 | CL | REG | ID\_unit | Prob | Stratum |
| 19 | 19 | 27 | 28 | 250 | 9 | 22 | 41 | 1616 | 2208 | 4 | 1 | 19 | 0.07042254 | 1 |
| 27 | 27 | 9 | 10 | 74 | 3 | 20 | 31 | 490 | 1751 | 6 | 2 | 27 | 0.07042254 | 2 |
| 195 | 195 | 17 | 17 | 128 | 6 | 25 | 45 | 1048 | 1606 | 35 | 2 | 53 | 0.07042254 | 2 |
| 208 | 208 | 8 | 8 | 60 | 6 | 19 | 39 | 390 | 723 | 37 | 2 | 66 | 0.07042254 | 2 |
| 57 | 57 | 31 | 33 | 220 | 10 | 27 | 57 | 1597 | 2653 | 11 | 3 | 79 | 0.07042254 | 3 |
| 72 | 72 | 8 | 8 | 41 | 7 | 15 | 41 | 349 | 765 | 13 | 3 | 94 | 0.07042254 | 3 |
| 93 | 93 | 14 | 14 | 86 | 6 | 17 | 41 | 574 | 1306 | 16 | 4 | 115 | 0.07042254 | 4 |
| 111 | 111 | 15 | 14 | 78 | 10 | 17 | 49 | 543 | 1622 | 19 | 4 | 133 | 0.07042254 | 4 |
| 124 | 124 | 21 | 19 | 122 | 10 | 15 | 49 | 917 | 2784 | 22 | 5 | 146 | 0.07042254 | 5 |
| 135 | 135 | 11 | 10 | 65 | 6 | 17 | 41 | 418 | 1135 | 24 | 5 | 157 | 0.07042254 | 5 |
| 176 | 176 | 13 | 13 | 83 | 5 | 19 | 41 | 623 | 1192 | 31 | 5 | 198 | 0.07042254 | 5 |
| 236 | 236 | 88 | 85 | 720 | 14 | 44 | 75 | 4758 | 8760 | 43 | 6 | 236 | 0.07042254 | 6 |
| 239 | 239 | 28 | 28 | 203 | 6 | 25 | 49 | 1524 | 2903 | 43 | 6 | 239 | 0.07042254 | 6 |
| 247 | 247 | 60 | 60 | 432 | 6 | 33 | 61 | 3070 | 6502 | 44 | 7 | 247 | 0.07042254 | 7 |
| 257 | 257 | 3 | 4 | 21 | 5 | 11 | 31 | 173 | 347 | 46 | 8 | 257 | 0.07042254 | 8 |
| 270 | 270 | 74 | 72 | 592 | 7 | 36 | 65 | 4777 | 7624 | 48 | 8 | 270 | 0.07042254 | 8 |

The frequency of each region in the sample, and the percentage of the region in the total data set are shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Frequency | % Sample | Percentage |
| 1 | 1 | 6.3 | 8.8 |
| 2 | 3 | 18.8 | 16.9 |
| 3 | 2 | 12.5 | 11.3 |
| 4 | 2 | 12.5 | 13.4 |
| 5 | 3 | 18.8 | 19.7 |
| 6 | 2 | 12.5 | 14.4 |
| 7 | 1 | 6.3 | 5.2 |
| 8 | 2 | 12.5 | 10.2 |

1. The means of the RMT85 variable of these four samples and the entire data set are shown in the table below. Given the results, it would appear that systematic sampling (where a member of each evenly sized group is drawn for the sample) is most representative of the population for the RMT85 variable.

|  |  |
| --- | --- |
| Sample | Mean |
| Full | 245 |
| RSWOR | 444 |
| Systematic | 245 |
| Systematic UP | 127 |
| Stratified | 198 |