

STAT 13, section 1, Winter 2012, UCLA Statistics

Solutions to Homework 2

**(HW2.1)** Suppose that in a certain population of married couples 35% of the husbands smoke, 15% of the wives smoke and in 10% of the couples both the husband and wife smoke. Is the smoking status of the husband independent of that of the wife? Why or why not?

No. Suppose they are independent, we should have

$$P(\text{Both the husband and wife smoke}) = P(\text{Husband smokes}) * P(\text{Wife smokes}).$$

$$\text{But } 0.1 \neq 0.35 * 0.15 = 0.0525,$$

So smoking status of the husband and that of wife are not independent.

**(HW2.2)** A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that:

$$p = 0.9$$

- o All 20 will be cured.

$$p^{20} = 0.9^{20} = 0.1216$$

- o All but 1 will be cured.

$$C_{20}^{19} * p^{19} * (1-p) = C_{20}^{19} * 0.9^{19} * 0.1 = 0.27$$

- o Exactly 17 will be cured.

$$C_{20}^{17} * p^{17} * (1-p)^3 = C_{20}^{17} * 0.9^{17} * 0.1^3 = 0.19$$

- o Exactly 80% will be cured.

$$20 * 0.8 = 16$$

$$C_{20}^{16} * p^{16} * (1-p)^4 = C_{20}^{16} * 0.9^{16} * 0.1^4 = 0.09$$

**(HW2.3)** Childhood lead poisoning is a public health concern in the US. In a certain population, one child in seven has a high blood lead level ( $>30 \text{ } \mu\text{g/dLi}$ ). Compute the following probabilities for a randomly chosen group of 16 children from this population:

$$p = \frac{1}{7}$$

- o  $P(\text{none have high blood lead})$

$$(1-p)^{16} = \left(\frac{6}{7}\right)^{16} = 0.085$$

- o  $P(\text{one has high blood lead})$

$$C_{16}^1 p^1 (1-p)^{15} = 16 \left(\frac{1}{7}\right) \left(\frac{6}{7}\right)^{15} = 0.226$$

- o  $P(\text{two have high blood lead})$

$$C_{16}^2 p^2 (1-p)^{14} = 120 \left(\frac{1}{7}\right)^2 \left(\frac{6}{7}\right)^{14} = 0.28$$

- o  $P(\text{three or more have high blood lead})$

$$1 - P(\text{none have high blood lead}) - P(\text{one has high blood lead}) - P(\text{two have high blood lead})$$

$$= 1 - 0.085 - 0.226 - 0.28 = 0.409$$

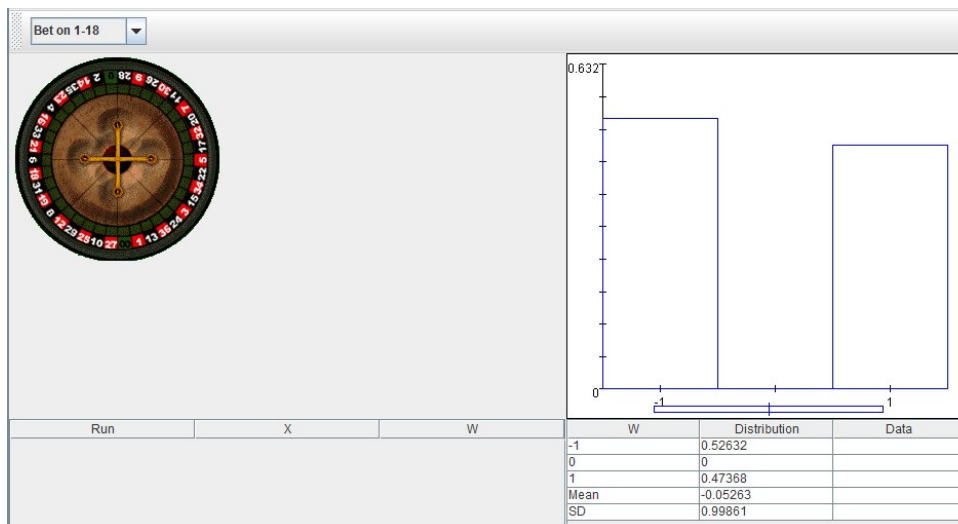
**(HW2.4)** Use the [SOCR Roulette Experiment](#) to design and run a simulation estimating the probability that a number between 19 and 36 turns up if we spin the [Roulette Wheel](#). Compute the exact probability of this event and list the sample-driven estimates of this event for samples of size 10, 100 and 1,000. What is your observation about these probability estimates?

Probability that a number  $x$  between 19 and 36 turns up is (18 numbers divided by the total 38 possible slots):

$$P(x \in [19, 36]) = \frac{18}{38} = 0.4737,$$

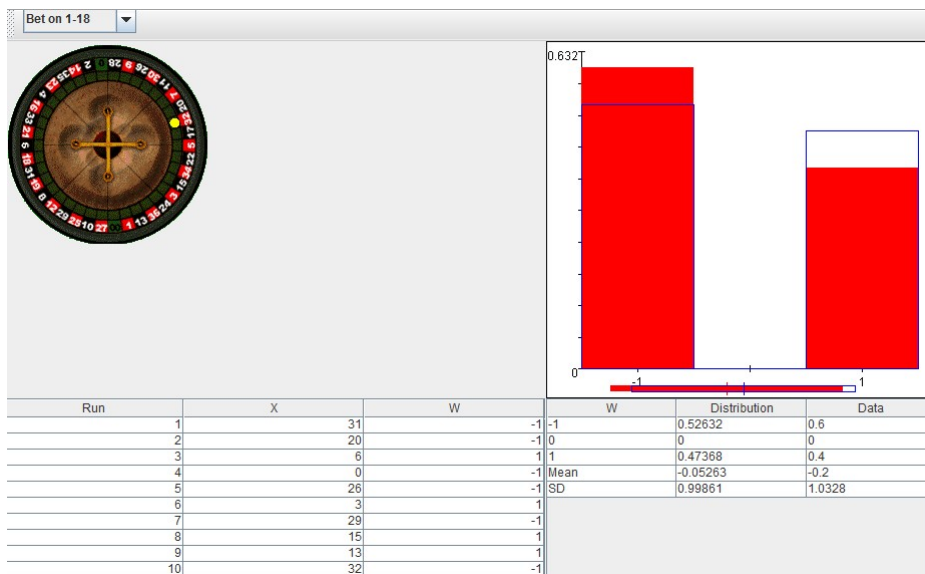
$$P(x \in \text{others}) = \frac{20}{38} = 0.5263.$$

In SOCR, we don't have the option to bet on 19 to 36. But we can do it by betting on 1 to 18 which will be of equal probability to bet on 19 to 36.

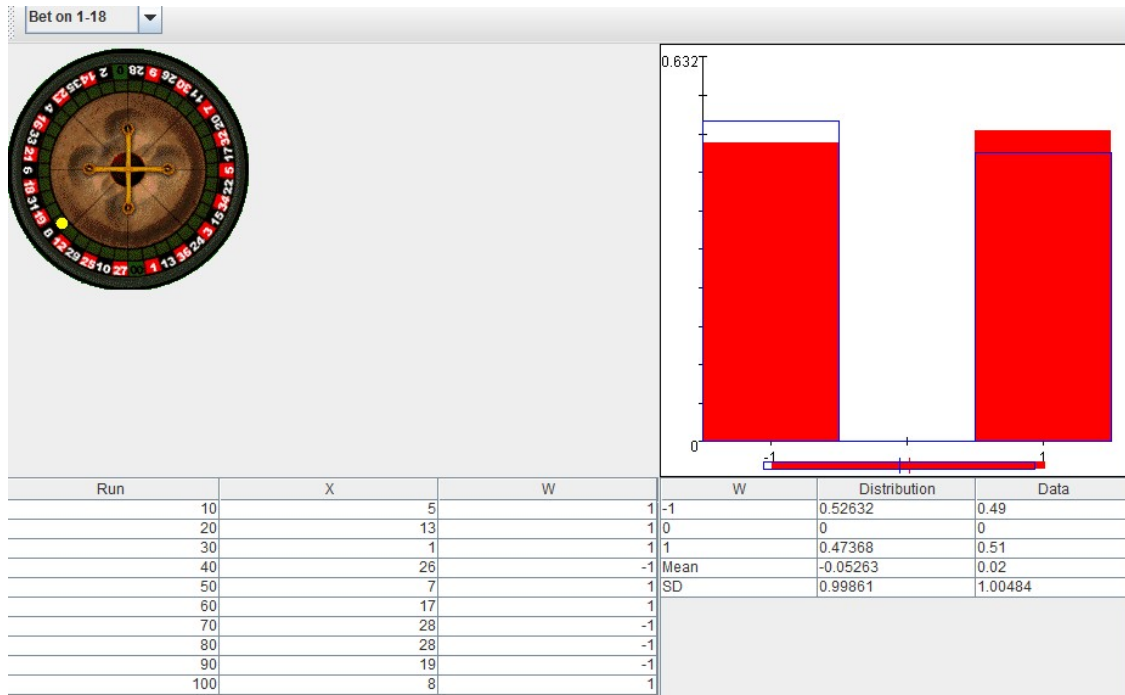


Virtual Experiments:

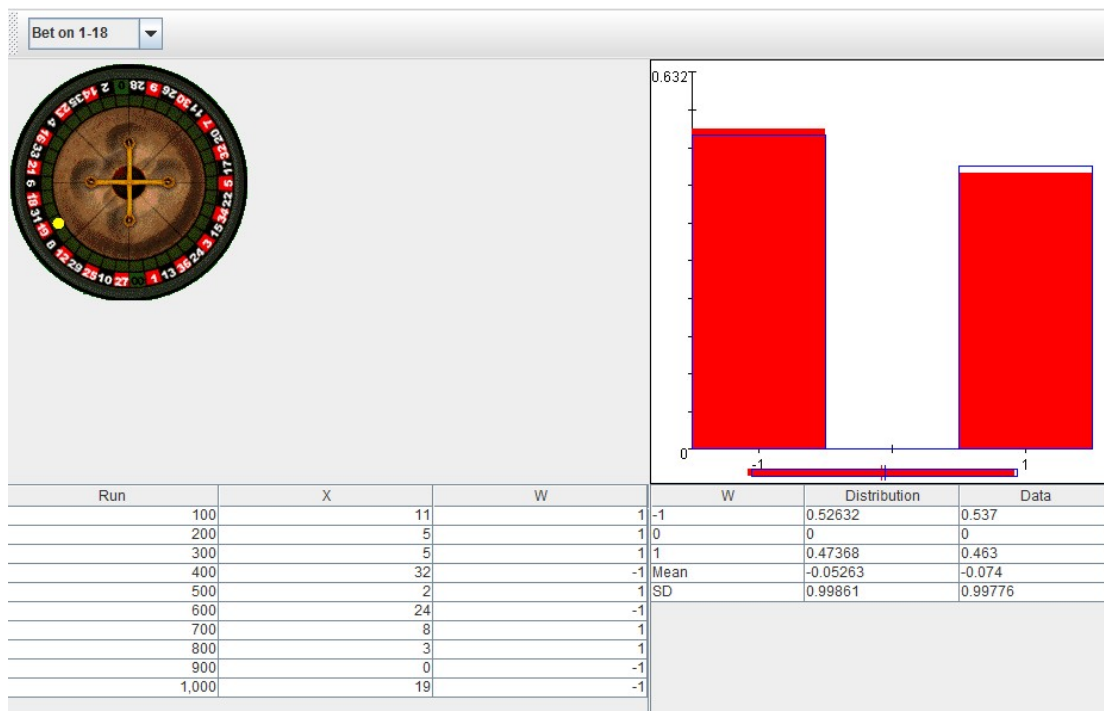
Samples of Size 10:



Samples of Size 100:



Samples of Size 1000:



Observation and Conclusion:

The more samples we have, the closer our empirical distribution (sample data distribution) will be to the theoretical distribution.

(HW2.5) Suppose that a long stretch of DNA has only Adenine (A), Thiamine (T), Cytosine (C) and Guanine (G) randomly occurring with the following probabilities 0.20, 0.3, 0.25, 0.25, respectively. The A, T, C and G nucleotides make up the core of the genetic code for any species. What is the probability that

- o A sample of 6 randomly chosen nucleotides contains 5 A's in a row?

$$P(AAAAA^*) + P(*AAAAA) + P(AAAAA) = 2 * 0.2^5 * 0.8 + 0.2^6 = 0.000576$$

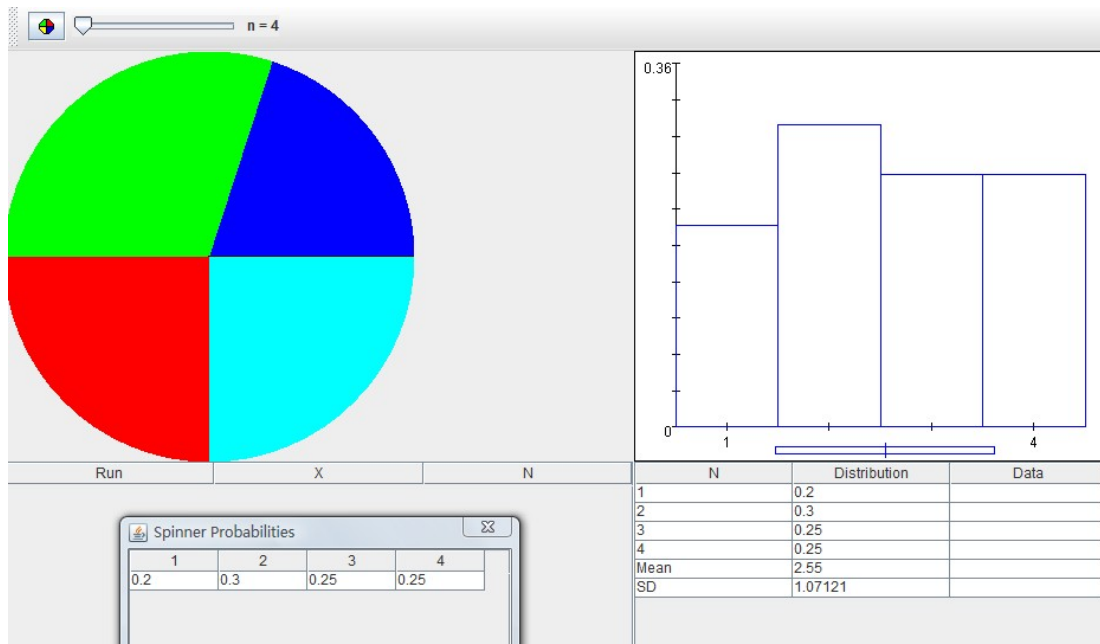
- o A random sample of 5 nucleotides has equal number of A's and T's?

$$P(\text{No A or T}) + P(\text{one A and one T}) + P(\text{two As and two Ts}) =$$

$$0.5^5 + C_5^1 C_4^1 * 0.2 * 0.3 * 0.5^3 + C_5^2 C_3^2 * 0.2^2 * 0.3^2 * 0.5 = 0.23525$$

- o Use the SOCR Spinner experiment (remember to set the right number of sectors, 4, and correct probabilities) to simulate the situations above and find empirical answers to the first 2 questions. Run 100 simulations and compare the empirical results of the applet with your theoretical calculations above.

Set up parameters:



Simulation for the first question: (The angle is starting from blue counter-clockwise)

A is blue (1); T is green (2); C is red (3); G is grey (4).

Here's the simulation result for 100 times.

1.0 348.05985110639523 4.0  
 2.0 184.65370701331753 3.0  
 3.0 14.658192576267357 1.0  
 4.0 312.31429615274 4.0  
 5.0 133.446778707856 2.0  
 6.0 189.6199583324132 3.0

7.0 292.62767567727894 4.0  
8.0 56.018499988677696 1.0  
9.0 14.508491793153201 1.0  
10.0 138.35344893982386 2.0  
11.0 139.86130654138876 2.0  
12.0 301.3651788088726 4.0  
  
13.0 1.8013448116206643 1.0  
14.0 92.07389916986591 2.0  
15.0 328.522692145794 4.0  
16.0 332.0031013841105 4.0  
17.0 354.7559268794021 4.0  
18.0 316.9800403614793 4.0  
  
19.0 319.49082238458027 4.0  
20.0 117.11946845041814 2.0  
21.0 46.12563719500921 1.0  
22.0 147.4111264340083 2.0  
23.0 35.176823201524094 1.0  
24.0 295.5654812011984 4.0  
  
25.0 64.04836355397961 1.0  
26.0 7.5286090658650195 1.0  
27.0 128.7824475649363 2.0  
28.0 177.08182521094903 2.0  
29.0 245.48200094698535 3.0  
30.0 264.5466589760282 3.0  
  
31.0 331.12438260928303 4.0  
32.0 296.7971838927876 4.0  
33.0 22.313974532360422 1.0  
34.0 318.1416174438929 4.0  
35.0 94.31298494129335 2.0  
36.0 326.3995276460821 4.0  
  
37.0 60.23389137335428 1.0

38.0 183.5220631602048 3.0  
39.0 320.98759954319183 4.0  
40.0 273.85625465314155 4.0  
41.0 80.49744247776552 2.0  
42.0 199.74277593280604 3.0  
  
43.0 268.47508063540187 3.0  
44.0 253.2860861334296 3.0  
45.0 335.998807432496 4.0  
46.0 30.19732583639633 1.0  
47.0 269.15190596014384 3.0  
48.0 277.56201471581255 4.0  
  
49.0 305.6079173554466 4.0  
50.0 338.8258076219337 4.0  
51.0 31.45454475218143 1.0  
52.0 238.4512024452695 3.0  
53.0 14.517798779748844 1.0  
54.0 63.42552264619948 1.0  
  
55.0 300.39675315502944 4.0  
56.0 132.57431558797066 2.0  
57.0 183.38660782353472 3.0  
58.0 106.5541927560178 2.0  
59.0 202.11631485091542 3.0  
60.0 226.193356108835 3.0  
  
61.0 275.34768504590585 4.0  
62.0 324.86673561543927 4.0  
63.0 273.8882882364023 4.0  
64.0 227.81349181583425 3.0  
65.0 167.50024519820008 2.0  
66.0 199.3282377887522 3.0  
  
67.0 349.20531579255754 4.0  
68.0 334.81120295070576 4.0

69.0 22.282325799856025 1.0  
70.0 271.597321275163 4.0  
71.0 246.94006166983422 3.0  
72.0 118.77727076788115 2.0  
  
73.0 334.16663841430363 4.0  
74.0 295.97565404773934 4.0  
75.0 124.2686211331336 2.0  
76.0 253.99009058008266 3.0  
77.0 198.40908250304213 3.0  
78.0 1.6708840220637011 1.0  
  
79.0 314.9480403012501 4.0  
80.0 354.41096363663644 4.0  
81.0 351.1186845831109 4.0  
82.0 226.72195043037132 3.0  
83.0 207.7049202429725 3.0  
84.0 106.81994326265306 2.0  
  
85.0 303.7068511389028 4.0  
86.0 92.78711228689582 2.0  
87.0 143.9062606222402 2.0  
88.0 71.18707544630949 1.0  
89.0 19.81290604675188 1.0  
90.0 279.09681714521344 4.0  
  
91.0 349.34782238557284 4.0  
92.0 44.100668053915044 1.0  
93.0 99.0917826850887 2.0  
94.0 44.22482417518133 1.0  
95.0 41.90920088838334 1.0  
96.0 267.05063412719636 3.0  
  
97.0 99.78900459482263 2.0  
98.0 315.93285588036724 4.0  
99.0 354.9889670671231 4.0

100.0 248.5915197160894 3.0

- *A sample of 6 randomly chosen nucleotides contains 5 A's in a row?*

For the first question, look at the simulation 6 a row and count qualified sets (five consecutive 1's). We should look from 1 to 96 simulations.

No consecutive 6 samples containing five consecutive 1's (five consecutive A's) in these 96 samples.

Theoretical:

$$P(AAAAA*) + P(*AAAAA) + P(AAAAAA) = 2 * 0.2^5 * 0.8 + 0.2^6 = 0.000576$$

Empirical (in 16 sets of 6 samples):

$$\frac{0}{16} = 0.$$

The theoretical probability is already very small here. So the empirical probability 0 is not surprising.

- *A random sample of 5 nucleotides has equal number of A's and T's?*

Look at simulations 5 a row and count qualified sets. (equal A's and Ts, say equal 1's and 2's). We can look from 1 to 100 simulations.

1.0 348.05985110639523 4.0  
2.0 184.65370701331753 3.0  
3.0 14.658192576267357 1.0  
4.0 312.31429615274 4.0  
5.0 133.446778707856 2.0

6.0 189.6199583324132 3.0  
7.0 292.62767567727894 4.0  
8.0 56.018499988677696 1.0  
9.0 14.508491793153201 1.0  
10.0 138.35344893982386 2.0

11.0 139.86130654138876 2.0  
12.0 301.3651788088726 4.0  
13.0 1.8013448116206643 1.0  
14.0 92.07389916986591 2.0  
15.0 328.522692145794 4.0



16.0 332.0031013841105 4.0  
17.0 354.7559268794021 4.0  
18.0 316.9800403614793 4.0  
19.0 319.49082238458027 4.0  
20.0 117.11946845041814 2.0  
  
21.0 46.12563719500921 1.0  
22.0 147.4111264340083 2.0  
23.0 35.176823201524094 1.0  
24.0 295.5654812011984 4.0  
25.0 64.04836355397961 1.0  
  
26.0 7.5286090658650195 1.0  
27.0 128.7824475649363 2.0  
28.0 177.08182521094903 2.0  
29.0 245.48200094698535 3.0  
30.0 264.5466589760282 3.0  
  
31.0 331.12438260928303 4.0  
32.0 296.7971838927876 4.0  
33.0 22.313974532360422 1.0  
34.0 318.1416174438929 4.0  
35.0 94.31298494129335 2.0  
  
36.0 326.3995276460821 4.0  
37.0 60.23389137335428 1.0  
38.0 183.5220631602048 3.0  
39.0 320.98759954319183 4.0  
40.0 273.85625465314155 4.0  
  
41.0 80.49744247776552 2.0  
42.0 199.74277593280604 3.0  
43.0 268.47508063540187 3.0  
44.0 253.2860861334296 3.0  
45.0 335.998807432496 4.0

46.0 30.19732583639633 1.0  
47.0 269.15190596014384 3.0  
48.0 277.56201471581255 4.0  
49.0 305.6079173554466 4.0  
50.0 338.8258076219337 4.0

51.0 31.45454475218143 1.0  
52.0 238.4512024452695 3.0  
53.0 14.517798779748844 1.0  
54.0 63.42552264619948 1.0  
55.0 300.39675315502944 4.0

56.0 132.57431558797066 2.0  
57.0 183.38660782353472 3.0  
58.0 106.5541927560178 2.0  
59.0 202.11631485091542 3.0  
60.0 226.193356108835 3.0

61.0 275.34768504590585 4.0  
62.0 324.86673561543927 4.0  
63.0 273.8882882364023 4.0  
64.0 227.81349181583425 3.0  
65.0 167.50024519820008 2.0

66.0 199.3282377887522 3.0  
67.0 349.20531579255754 4.0  
68.0 334.81120295070576 4.0  
69.0 22.282325799856025 1.0  
70.0 271.597321275163 4.0

71.0 246.94006166983422 3.0  
72.0 118.77727076788115 2.0  
73.0 334.16663841430363 4.0  
74.0 295.97565404773934 4.0  
75.0 124.2686211331336 2.0

76.0 253.99009058008266 3.0  
77.0 198.40908250304213 3.0  
78.0 1.6708840220637011 1.0  
79.0 314.9480403012501 4.0  
80.0 354.41096363663644 4.0

81.0 351.1186845831109 4.0  
82.0 226.72195043037132 3.0  
83.0 207.7049202429725 3.0  
84.0 106.81994326265306 2.0  
85.0 303.7068511389028 4.0

86.0 92.78711228689582 2.0  
87.0 143.9062606222402 2.0  
88.0 71.18707544630949 1.0  
89.0 19.81290604675188 1.0  
90.0 279.09681714521344 4.0

91.0 349.34782238557284 4.0  
92.0 44.100668053915044 1.0  
93.0 99.0917826850887 2.0  
94.0 44.22482417518133 1.0  
95.0 41.90920088838334 1.0

96.0 267.05063412719636 3.0  
97.0 99.78900459482263 2.0  
98.0 315.93285588036724 4.0  
99.0 354.9889670671231 4.0  
100.0 248.5915197160894 3.0

Theoretical:

$P(\text{No A or T}) + P(\text{one A and one T}) + P(\text{two As and two Ts}) =$

$$0.5^5 + C_5^1 C_4^1 * 0.2 * 0.3 * 0.5^3 + C_5^2 C_3^2 * 0.2^2 * 0.3^2 * 0.5 = 0.23525$$

Empirical (in 20 sets of 5 samples):

[http://www.stat.ucla.edu/~dinov/courses\\_students.dir/12/Winter/STAT13.1.dir](http://www.stat.ucla.edu/~dinov/courses_students.dir/12/Winter/STAT13.1.dir)

$$\frac{3}{20} = 0.15 .$$

The results are not very close, but to verify our theoretical distribution we need more simulations. We only have 20 here. (5 as a set)