Spreadsheets in Education (eJSiE)

Volume 6 | Issue 3 Article 3

10-20-2013

Sampling with & without Replacement: Urn problem modeled with Geogebra

Gunhan Caglayan
Florida International University, gcaglayan@njcu.edu

Sampling with & without Replacement: Urn problem modeled with Geogebra

Abstract
Using spreadsheets, students could perform probability experiments, illustrate the outcomes, experience, understand, and represent mathematics with motivation in an inspirational manner. Although GeoGebra Dynamic Software has been mostly used with geometry and algebra explorations, its recently improved statistics feature in Version 4 provides students and teachers with more opportunities for probability and statistics investigations. The lesson presented in this article emphasizes the use of spreadsheets to record the outcomes of probability experiments as a method for delivering mathematical ideas, which support the teaching of sampling with / without replacement in the context of urn problem.

Keywords
sampling; urn problem; spreadsheets; binomial distribution; hypergeometric distribution

Sampling with & without Replacement: Urn problem modeled with Geogebra

Gunhan Caglayan
Florida International University
gcaglaya@fiu.edu

Abstract

Using spreadsheets, students could perform probability experiments, illustrate the outcomes, experience, understand, and represent mathematics with motivation in an inspirational manner. Although GeoGebra Dynamic Software (geogebra.org) has been mostly used with geometry and algebra explorations, its recently improved statistics feature in Version 4 provides students and teachers with more opportunities for probability and statistics investigations. I present below a lesson outline based on my college teaching experiences from an introductory probability course offered to math and math education majors in their third year. This lesson emphasizes use of spreadsheets to record the outcomes of probability experiments as a method for delivering mathematical ideas, which stands as an effective way of exploring mathematics. These notes support the teaching of sampling with / without replacement in the context of urn problem. The mathematical experience necessitates active participation of the teacher and the students.

Keywords: sampling; urn problem; spreadsheets; binomial distribution; hypergeometric distribution

1. Sampling with replacement

This activity is based on Example 5b, from *A First Course in Probability* text by *Sheldon Ross*, which was also the textbook I used with my students. Example 5b (p. 34) actually models sampling without replacement with the urn problem. Because it is easier to model it on GeoGebra, I start the discussion with sampling with replacement.

Experiment: Three balls are randomly drawn (with replacement) from an urn containing six white and five black balls. X is the random variable giving the number of white balls drawn. Events of interest are $X = 0$, $X = 1$, $X = 2$, and $X = 3$. Perform this experiment 1000 times. Record the outcomes. Create a table (including a "total" column). Compare and comment on the experimental vs. theoretical probabilities.

Prior to engaging in a GeoGebra simulation, I first asked my students to make guesses about the most/least likely outcomes. Almost all students conjectured that the events $X = 2$ or $X = 3$ (i.e., samples containing two or three white balls) would be more likely to occur than the others because there were more white balls than black balls in the urn. Before modeling this experiment on GeoGebra, we selected the Spreadsheet perspective. The conventions we used were 1 and 2, as suggested by students, respectively standing for the white and the black balls. We induced an urn containing six white and five black balls by typing bag = {1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2} in the input bar. On the spreadsheet, each row represents a single trial and columns A, B, C respectively stand for the first, the second, and the third ball drawn from the urn. We used the RandomElement built-in function to draw these balls:

A1 = RandomElement[bag]

B1 = RandomElement[bag]

C1 = RandomElement[bag]

We used the CountIf function in Column D to count the number of white balls in the first trial:

D1 = CountIf[x == 1, A1:C1]

Selecting cells A1, B1, C1, D1 and dragging the cursor down up to the 1000th row enabled us to perform the experiment 1000 times very quickly. To record the outcomes, we selected a table space (for a 4 by 5 table) far from the columns A, B, C, D. We labeled the cells of the first column of the table as Event, Frequency, Experimental Probability, and Theoretical Probability. The cells of the first row of the table were labeled as "X = 0," "X = 1," "X = 2," "X = 3," and "Total." Finally, for each event respectively, we used the CountIf function to count the number of white balls overall for the 1000 trials:

G2 = CountIf[x == 0, D1:D1000]

H2 = CountIf[x == 1, D1:D1000]

I2 = CountIf[x == 2, D1:D1000]

J2 = CountIf[x == 3, D1:D1000]

The rationale for including a Total column was for normalization purpose and to make sure that the sum of the frequencies would yield the number of trials, namely 1000. We used the Sum function to add the frequencies:

K2 = Sum[G2:J2]

To compute the experimental probability for the first event, we typed:

G3 = G2 / K2

The $ symbol was used to fix the value of K2. By dragging the cursor to the right, we calculated the experimental probabilities for the other events as well. As for the theoretical probabilities, we found:

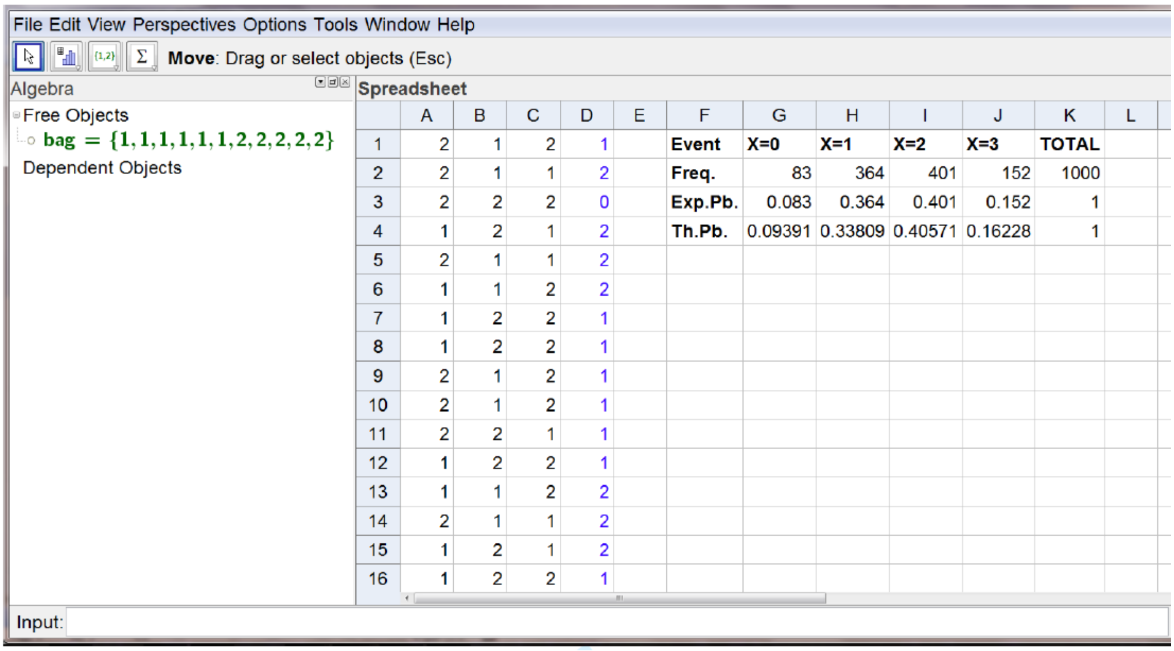
$$P(X=0)=\frac{5}{11}\frac{5}{11}\frac{5}{11}=\frac{125}{1331}\approx 9.39\%$$

$$P(X=1)=\frac{6}{11}\frac{5}{11}\frac{5}{11}+\frac{5}{11}\frac{6}{11}\frac{5}{11}+\frac{5}{11}\frac{5}{11}\frac{6}{11}=\frac{450}{1331}\approx 33.81\%$$

$$P(X=2)=\frac{6}{11}\frac{6}{11}\frac{5}{11}+\frac{6}{11}\frac{5}{11}\frac{6}{11}+\frac{5}{11}\frac{6}{11}\frac{6}{11}=\frac{540}{1331}\approx 40.57\%$$

$$P(X=3)=\frac{6}{11}\frac{6}{11}\frac{6}{11}=\frac{216}{1331}\approx 16.23\%$$

We entered these values in cells G4 through J4 for comparison purposes. Figure 1 depicts the above discussion.



	A	B	C	D	E	F	G	H	I	J	K	L
1	2	1	2	1		Event	X=0	X=1	X=2	X=3	TOTAL	
2	2	1	1	2		Freq.	83	364	401	152	1000	
3	2	2	2	0		Exp.Pb.	0.083	0.364	0.401	0.152	1	
4	1	2	1	2		Th.Pb.	0.09391	0.33809	0.40571	0.16228	1	
5	2	1	1	2								
6	1	1	2	2								
7	1	2	2	1								
8	1	2	2	1								
9	2	1	2	1								
10	2	1	2	1								
11	2	2	1	1								
12	1	2	2	1								
13	1	1	2	2								
14	2	1	1	2								
15	1	2	1	2								
16	1	2	2	1								

Figure 1: Sampling with Replacement Experiment on GeoGebra Display

Performing the experiment 5,000 times gave us seemingly better experimental results:

F	G	H	I	J	K
Event	X=0	X=1	X=2	X=3	TOTAL
Freq.	480	1676	1996	848	5000
Exp.Pb.	0.096	0.3352	0.3992	0.1696	1
Th.Pb.	0.09391	0.33809	0.40571	0.16228	1

Figure 2: Sampling with Replacement Experiment Performed 5000 Times

2. Sampling without replacement

Experiment: Three balls are randomly drawn (without replacement) from an urn containing six white and five black balls. X is the random variable giving the number of white balls drawn. Events of interest are $X = 0$, $X = 1$, $X = 2$, and $X = 3$. Perform this experiment 1000 times. Record the outcomes. Create a table (including a "total" column). Compare and comment on the experimental vs. theoretical probabilities.

As was the case for the previous experiment, prior to engaging in a GeoGebra simulation, I asked students to make guesses about the most/least likely outcomes. This time, most students conjectured that the events $X = 1$ or $X = 2$ (i.e., samples containing one or two white balls) would be more likely to occur than the others. With the same conventions and the same urn, we used the RandomElement built-in function to draw the first ball:

A1 = RandomElement[bag]

Because the color of the second ball depends on the first ball drawn, we could not use the same function for B1. A slight modification based on conditioning (using the If, First, and Last built-in functions) was necessary:

B1 = If[A1 == 1, RandomElement[Last[bag, 10]], RandomElement[First[bag, 10]]]

Explanation for this command: If the first drawn ball is 1 (namely white), then draw a ball from the last 10 elements of the bag (namely from the bag containing 5 white 5 black balls). Or else (namely if the first drawn ball is 2, i.e. black), draw a ball from the first 10 elements of the bag (namely from the bag containing 6 white 4 black balls). The function that is used for C1 follows a similar reasoning (using the If, First, Last, and Take built-in functions) that involves more if-then loops:

C1 = If[A1 == 1, If[B1 == 1, RandomElement[Last[bag, 9]], RandomElement[Take[bag, 2, 10]]], If[B1 == 1, RandomElement[Take[bag, 2, 10]], RandomElement[First[bag, 9]]]]

Explanation for this command

<u>Case One</u>: The first drawn ball is 1 (namely white). If the second drawn ball is also 1, then draw a ball from the last 9 elements of the bag (namely from the bag containing 4 white 5 black balls). Or else (namely if the second drawn ball is 2, i.e. black), draw a ball from the elements from position 2 to 10 of the initial bag (namely from the bag containing 5 white 4 black balls).

<u>Case Two</u>: The first drawn ball is 2 (namely black). If the second drawn ball is 1 (namely white), then draw a ball from the elements from position 2 to 10 of the initial bag (namely from the bag containing 5 white 4 black balls). Or else (namely if the second drawn ball is 2, i.e. black), then draw a ball from the first 9 elements of the bag (namely from the bag containing 6 white 3 black balls). The remaining procedure was the same as the sampling without replacement experiment.

As for the theoretical probabilities, we found:

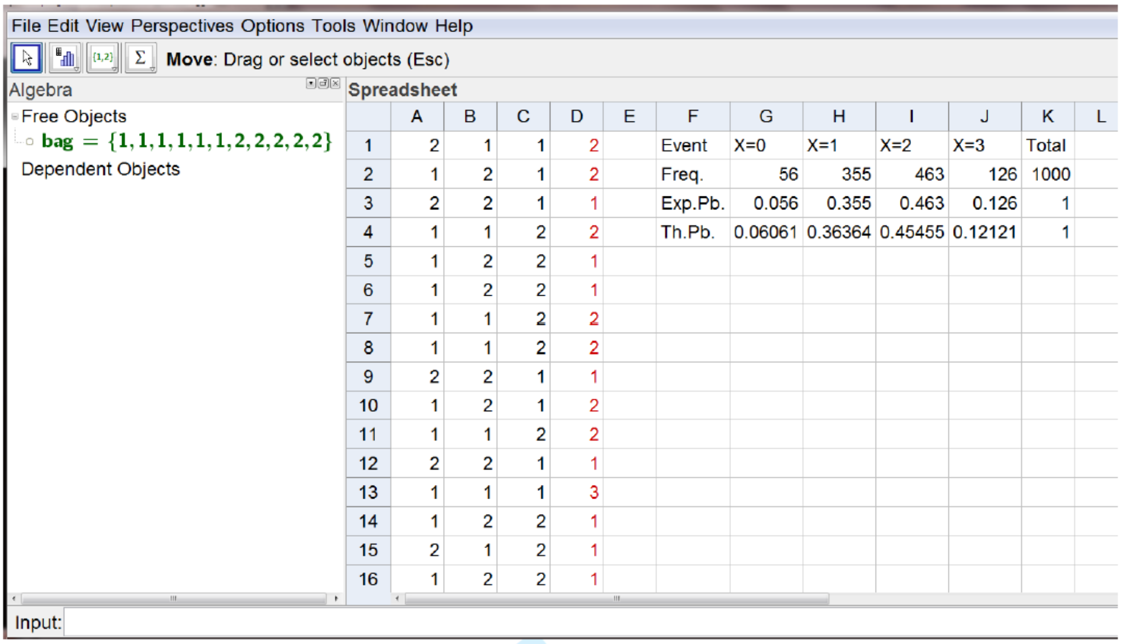
$$P(X=0)=\frac{5}{11}\frac{4}{10}\frac{3}{9}=\frac{60}{990}\approx 6.06\%$$

$$P(X=1)=\frac{6}{11}\frac{5}{10}\frac{4}{9}+\frac{5}{11}\frac{6}{10}\frac{4}{9}+\frac{5}{11}\frac{4}{10}\frac{6}{9}=\frac{360}{990}\approx 36.36\%$$

$$P(X=2)=\frac{6}{11}\frac{5}{10}\frac{5}{9}+\frac{6}{11}\frac{5}{10}\frac{5}{9}+\frac{5}{11}\frac{6}{10}\frac{5}{9}=\frac{450}{990}\approx 45.45\%$$

$$P(X=3)=\frac{6}{11}\frac{5}{10}\frac{4}{9}=\frac{120}{990}\approx 12.12\%$$

We entered these values in cells G4 through J4 for comparison purposes as before. Figure 3 depicts the discussion above.



	A	B	C	D	E	F	G	H	I	J	K	L
1	2	1	1	2		Event	X=0	X=1	X=2	X=3	Total	
2	1	2	1	2		Freq.	56	355	463	126	1000	
3	2	2	1	1		Exp.Pb.	0.056	0.355	0.463	0.126	1	
4	1	1	2	2		Th.Pb.	0.06061	0.36364	0.45455	0.12121	1	
5	1	2	2	1								
6	1	2	2	1								
7	1	1	2	2								
8	1	1	2	2								
9	2	2	1	1								
10	1	2	1	2								
11	1	1	2	2								
12	2	2	1	1								
13	1	1	1	3								
14	1	2	2	1								
15	2	1	2	1								
16	1	2	2	1								

Figure 3: Sampling without Replacement Experiment on GeoGebra Display

3. Probability Calculator

The Probability Calculator feature in GeoGebra 4.0 helped us visualize and compare our experimental results with the theoretical ones. For this purpose, we used the binomial probability distribution $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ for the sampling with replacement experiment for which the parameters were $n = 3$, $p = 6/11$, and $k = 0, 1, 2, 3$. Using the probability calculator in GeoGebra, we obtained the theoretical probability distribution by entering 3 for n and 6/11 for p (i.e., the probability of "success") as depicted in Figure 4.

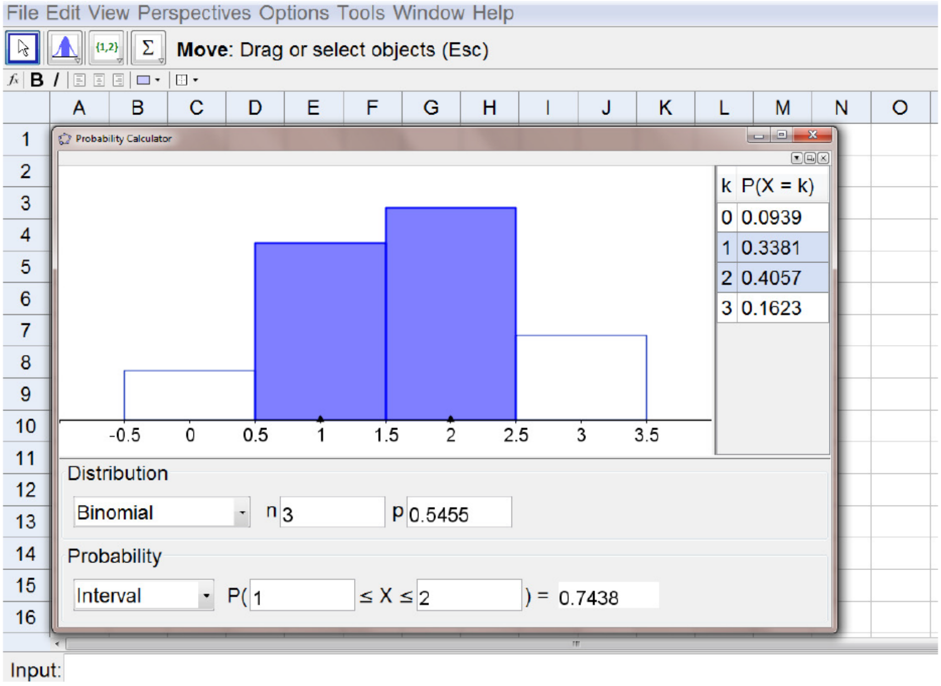


Figure 4: Binomial Distribution with GeoGebra's Probability Calculator

In a similar manner, the Probability Calculator helped us visualize the hypergeometric probability distribution $P(X=k)=\dfrac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$ for the sampling without replacement experiment for which the parameters were N = 11, n = 6, m = 3, and k = 0, 1, 2, 3. Using the probability calculator in GeoGebra, we obtained the theoretical probability distribution as depicted in Figure 5.

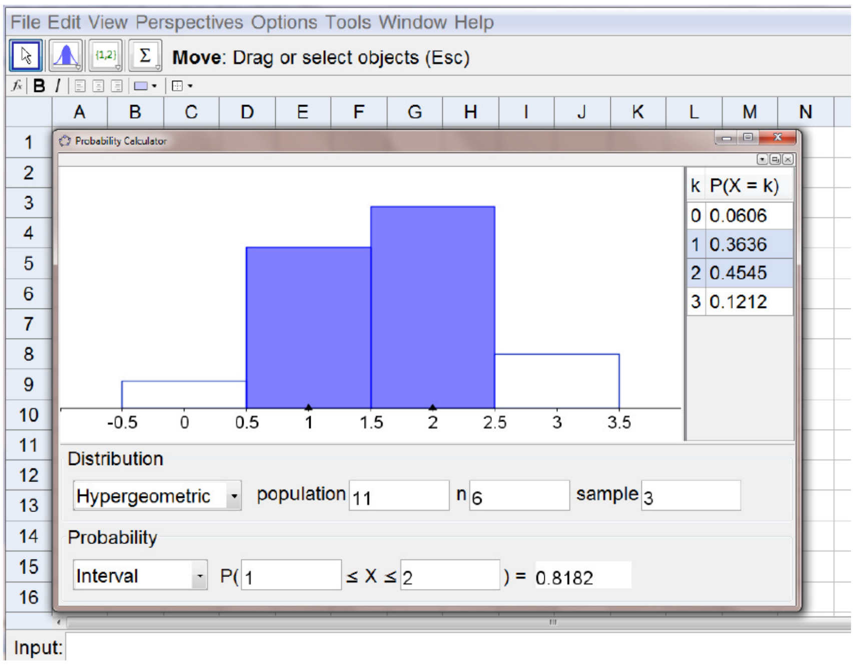


Figure 5: Hypergeometric Distribution with GeoGebra's Probability Calculator

We also explored extension problems, such as the exploration of various configurations (by keeping the total number of balls fixed). Figure 6 depicts the sampling with replacement (binomial distribution) scenarios with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 white balls.

p	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	total
0.09091	**0.75131**	0.22539	0.02254	0.00075	1
0.18182	**0.54771**	0.36514	0.08114	0.00601	1
0.27273	0.38467	**0.43276**	0.16228	0.02029	1
0.36364	0.2577	**0.44177**	0.25244	0.04808	1
0.45455	0.16228	**0.40571**	0.33809	0.09391	1
0.54545	0.09391	0.33809	**0.40571**	0.16228	1
0.63636	0.04808	0.25244	**0.44177**	0.2577	1
0.72727	0.02029	0.16228	**0.43276**	0.38467	1
0.81818	0.00601	0.08114	0.36514	**0.54771**	1
0.90909	0.00075	0.02254	0.22539	**0.75131**	1

Figure 6: Sampling with Replacement Scenarios with 1 through 10 white balls.

Some students suggested we graph these probabilities using GeoGebra's Graphics view (Figure 7). Figures 6 and 7 reveal that for the six out of ten situations, obtaining one or two white balls are the more likely scenarios. Figure 7 also reveals the fact that the probabilities are symmetric with respect to the $p = 0.5$ axis.

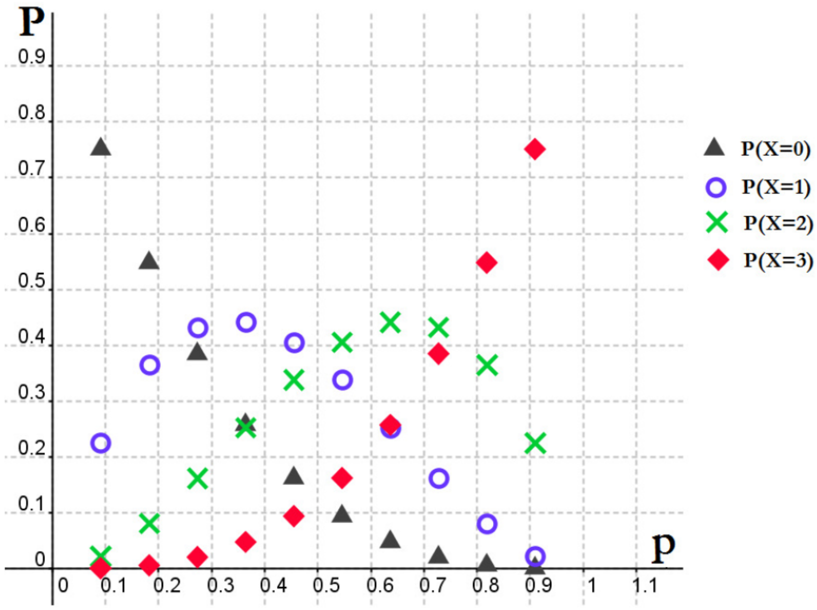


Figure 7: Scatter-Plots Illustrating Sampling with Replacement Scenarios

In a similar manner, we obtained the following table and scatter-plot for the sampling without replacement (hypergeometric distribution) scenarios.

n	P(X=0)	P(X=1)	P(X=2)	P(X=3)	total
1	**0.72727**	0.27273	0	0	1
2	**0.50909**	0.43636	0.05455	0	1
3	0.33939	**0.50909**	0.14545	0.00606	1
4	0.21212	**0.50909**	0.25455	0.02424	1
5	0.12121	**0.45455**	0.36364	0.06061	1
6	0.06061	0.36364	**0.45455**	0.12121	1
7	0.02424	0.25455	**0.50909**	0.21212	1
8	0.00606	0.14545	**0.50909**	0.33939	1
9	0	0.05455	0.43636	**0.50909**	1
10	0	0	0.27273	**0.72727**	1

Figure 8: Sampling without Replacement Scenarios with 1 through 10 white balls

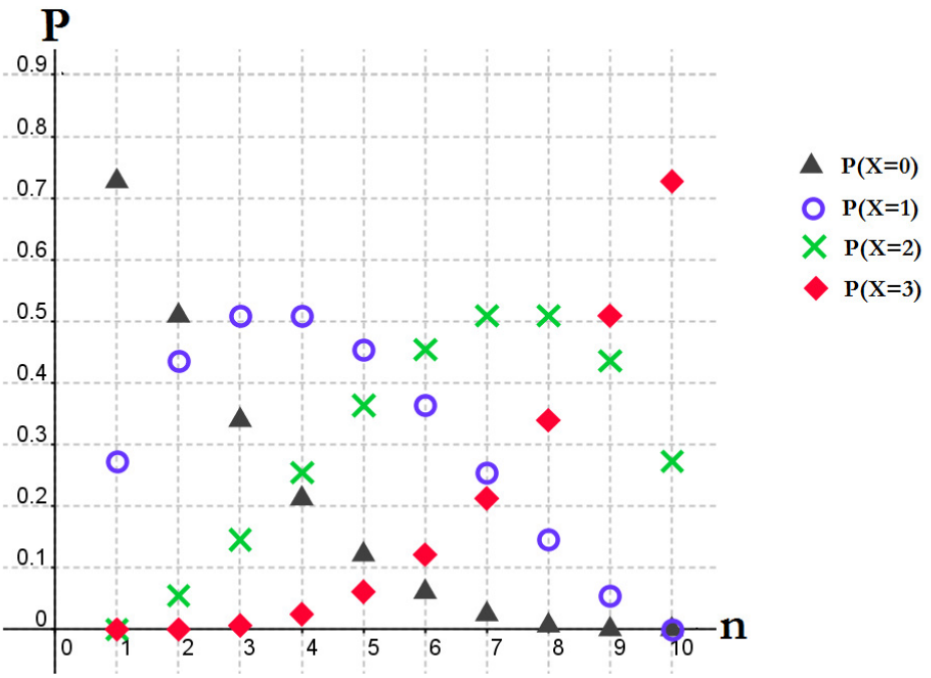


Figure 9: Scatter-Plots Illustrating Sampling without Replacement Scenarios

4. Reflections

Before exploring the probability experiments with the help of dynamic software like GeoGebra, it is essential that students have a good grasp of basic probability ideas and have already performed easier probability experiments using concrete objects such as coins, dice, cards, and spinners. I used this GeoGebra based lesson with math and math education majors (third-year university students) in the investigation of urn problem (sampling with and without replacement) situations. The lesson presented in this article depicts an advanced topic in probability, which requires a profound understanding of the prerequisite material, such as basic principles of combinatorial analysis, chance, randomness, sample spaces and events, axioms of probability theory.

It is worth noting that the outcomes of the experiments with the urn problem did not seem to surprise my students – their intuition was confirmed by the theory. Prior to these experiments, we also explored various simulations on GeoGebra involving coins, dice, and spinners, some of which were counterintuitive examples triggering students' probabilistic thinking and reasoning in a surprising – yet productive manner. For instance, in the experiment of tossing two fair coins, where X denotes the number of heads that appear, some students might intuitively suggest that all $X = 0$, $X = 1$, $X = 2$ are equally likely events before simulation (The most likely event is actually $X = 1$ with probability 0.5). Similarly, in the experiment of tossing two six-sided fair dice, where X designates the sum, some students might intuitively suggest the events $X = 6$, $X = 7$ and $X = 8$ to be more likely than the others based on an incorrect description of the sample space with twenty-one outcomes (Figure 10) instead of thirty-six (The most likely event is actually $X = 7$ with probability 6/36 = 1/6).

$(1,1) \to X = 2$	$(1,5),(2,4),(3,3) \to X = 6$	$(3,6),(4,5) \to X = 9$
$(1,2) \to X = 3$	$(1,6),(2,5),(3,4) \to X = 7$	$(4,6),(5,5) \to X = 10$
$(1,3),(2,2) \to X = 4$	$(2,6),(3,5),(4,4) \to X = 8$	$(5,6) \to X = 11$
$(1,4),(2,3) \to X = 5$		$(6,6) \to X = 12$

Figure 10: Incorrect Modeling of the Sample Space

Simulations are powerful problem-solving instruments that can help students visualize and make better sense of probability experiments. They also provide an effortless way of comparing initial predictions and experimental data along with algebraically derived theoretical results. GeoGebra's spreadsheets tool could be used to help students visualize popular experiments with coins, dice, spinners, cards, and marbles-in-urns as well as real-world situations such as quality control, casino games, insurance premiums, weather, and sports. When performing these experiments, it is essential that the facilitator in charge help students realize the impact of increasing the number of simulations on getting "seemingly better results," and its connection to the theoretical (limit-based) definition of probability $P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$ as the limiting frequency of the event E under consideration.

References

1. GeoGebra 4.0 (geogebra.org).
2. Ross, S. (2008). *A First Course in Probability* (8th Edition). Pearson.