



DSFA
Spring 2021

Lecture 15

Sampling

Announcements

- Wednesday afternoon lab online
 - Prelim 1 online; Zoom link will be posted in Canvas this afternoon; more instructions forthcoming.
-

When poll is active, respond at pollev.com/dsfa

Text **DSFA** to **22333** once to join

What's the probability of getting at least one 6 in 20 rolls of a 6-sided die?

At least 50% but less than 80%

At least 80% but less than 90%

At least 90% but less than 95%

At least 95% but less than 99%

At least 99%



Which do not return a random number from 1 to 6 (inclusive), with each number equally likely?

```
np.random.choice(np.arange(6))
```

```
np.random.choice(np.arange(6)+1)
```

```
np.random.choice(np.arange(1,6+1))
```

```
np.random.choice(np.arange(1,3+1)) +  
np.random.choice(np.arange(1,3+  
1))
```



Probability and Simulation

New York State presidential polling

Poll source	Date(s) administered	Sample size ^[b]	Margin of error	Donald Trump Republican	Joe Biden Democratic	Jo Jorgensen Libertarian	Howie Hawkins Green	Other	Undecided
SurveyMonkey/Axios	Oct 20 – Nov 2, 2020	6,548 (LV)	± 2%	35% ^[c]	63%	–	–	–	–
Research Co.	Oct 31 – Nov 1, 2020	450 (LV)	± 4.6%	34%	64%	-	-	2% ^[d]	4%
SurveyMonkey/Axios	Oct 1–28, 2020	10,220 (LV)	–	34%	63%	-	-	–	–
Swayable	Oct 23–26, 2020	495 (LV)	± 5.8%	33%	65%	1%	1%	–	–
SurveyMonkey/Axios	Sep 1–30, 2020	10,007 (LV)	–	34%	64%	-	-	–	2%
Siena College	Sep 27–29, 2020	504 (LV)	± 4.4%	29%	61%	0%	1%	2% ^[e]	7%
SurveyMonkey/Axios	Aug 1–31, 2020	9,969 (LV)	–	34%	64%	-	-	–	2%
Public Policy Polling	Aug 20–22, 2020	1,029 (V)	± 3.1%	32%	63%	-	-	–	5%
SurveyMonkey/Axios	Jul 1–31, 2020	10,280 (LV)	–	34%	63%	-	-	–	2%
SurveyMonkey/Axios	Jun 8–30, 2020	4,555 (LV)	–	33%	65%	-	-	–	2%
Siena College 	Jun 23–25, 2020	806 (RV)	± 3.9%	32%	57%	-	-	–	10%
Siena College 	May 17–21, 2020	767 (RV)	± 3.7%	32%	57%	-	-	–	11%
Quinnipiac University	Apr 30 – May 4, 2020	915 (RV)	± 3.2%	32%	55%	-	-	5% ^[f]	8%
Siena College 	Apr 19–23, 2020	803 (RV)	± 3.7%	29%	65%	-	-	–	6%
Siena College 	Mar 22–26, 2020	566 (RV)	± 4.5%	33%	58%	-	-	–	10%
Siena College 	Feb 16–20, 2020	658 (RV)	± 4.5%	36%	55%	-	-	–	5%

New York State presidential results

2020 United States presidential election in New York ^[38]				
Party	Candidate	Votes	%	±%
Democratic	<i>Joe Biden</i> <i>Kamala Harris</i>	4,844,975	56.37	-0.35%
Working Families	<i>Joe Biden</i> <i>Kamala Harris</i>	386,010	4.49	+2.68%
Total	Joe Biden Kamala Harris	5,230,985	60.86	+1.85%
Republican	<i>Donald Trump</i> <i>Mike Pence</i>	2,949,141	34.31	+1.58%
Conservative	<i>Donald Trump</i> <i>Mike Pence</i>	295,657	3.44	-0.35%
Total	Donald Trump Mike Pence	3,244,798	37.75	+1.23%
Libertarian	Jo Jorgensen Spike Cohen	60,234	0.70	-0.04%
Green	Howie Hawkins Angela Walker	32,753	0.38	-1.02%
Independence	Brock Pierce Karla Ballard	22,587	0.26	-1.28%
Write-in		3,469	0.04	-0.75%
	Total votes	8,594,826	100.00%	+11.31%

Sampling

Sampling

Observe some *individuals* from a *population*

- a. Examine 10 rolls of a d6 (six-sided die)
 - b. Coat color of the first 20 people who walk through door
 - c. Survey 1000 students living in campus dorms, where every student on campus is equally likely to be chosen, and ask them who they would vote for for president
-

Sampling

- Deterministic sample:
 - Sampling scheme doesn't involve chance
- Probability (random) sample:
 - Before the sample is drawn, you have to know the selection probability of every group of people in the population
 - Not all individuals have to have equal chance of being selected

(Demo)

When poll is active, respond at pollev.com/dsfa

Text **DSFA** to **22333** once to join

Which of these is a deterministic sample?

10 rolls of a 6-sided die

Coat color of first 20 people
entering the room

Survey of 100 students living
in dorms, in which student is
equally likely to be chosen



Sample of Convenience

- Example: sample consists of whoever walks by
 - Just because you think you're sampling "at random", doesn't mean you are. If you can't figure out ahead of time
 - what's the population
 - what's the chance of selection, for each group in the population
- then you don't have a random sample
-

**Does sample look like
population?**

(Demo)

Large Random Samples

If the sample size is large,

then the **empirical distribution** of a **uniform random** sample

resembles the **population distribution**,

with high probability.

Distribution

- A **distribution** is a description of the likelihood of *events*
- **Empirical** distribution:
 - Experimental: made from observations
 - Proportion of each event in sample

vs.

- **Probability** distribution:
 - Theoretical: made from mathematics
 - Probability of each event
-

Law of Large Numbers

If an experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event

Sometimes called *Law of Averages*
