DSC 40A

Theoretical Foundations of Data Science I

Random Sampling

Agenda

- Conditional probability continued
- Sampling with and without replacement

Question

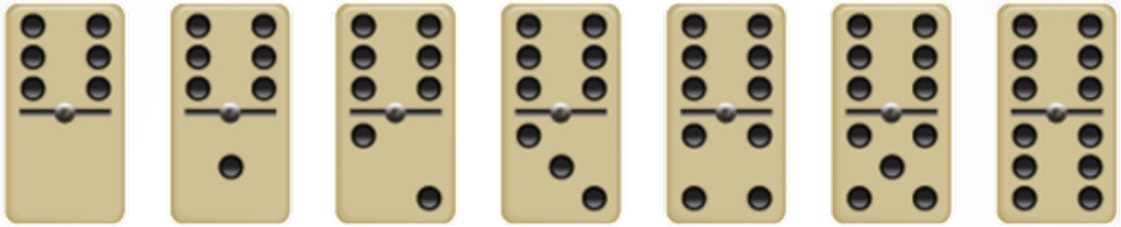
Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!
If the direct link doesn't work, click the "Lecture Questions" link in the top right corner of dsc40a.com.

Conditional probability continued

Dominoes

Question 3: Now you pick a random tile from the set and uncover only one side, revealing that it has 6 dots. What is the probability that this tile is a double, with 6 on both sides?



Try it out in code!

Five Thirty Eight Riddler Express Solution: https://fivethirtyeight.com/features/can-you-find-the-best-dungeons-dragons-strategy/

Conditional probabilities: Simpson's Paradox

	Treatment A	Treatment B
Small kidney stones	81 successes / 87 (93%)	234 successes / 270 (87%)
Large kidney stones	192 successes / 263 (73%)	55 successes / 80 (69%)
Combined	273 successes / 350 (78%)	289 successes / 350 (83%)

Which treatment is better?

40% A. Treatment A for all cases.

14% B. Treatment B for all cases.

13% C. A for small and B for large.

30% D. A for large and B for small.

C. R. Charig, D. R. Webb, S. R. Payne, J. E. Wickham (29 March 1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". Br Med J (Clin Res Ed) 292 (6524): 879–882. doi:10.1136/bmj.292.6524.879. PMC 1339981. PMID 3083922. cf. Wikipedia "Simpson's Paradox"

Conditional probabilities: Simpson's Paradox

	Treatment A	Treatment B
Small kidney stones	81 successes / 87 (93%)	234 successes / 270 (87%)
Large kidney stones	192 successes / 263 (73%)	55 successes / 80 (69%)
Combined	273 successes / 350 (78%)	289 successes / 350 (83%)

Simpson's Paradox

"When the less effective treatment is applied more frequently to easier cases, it can appear to be a more effective treatment."

C. R. Charig, D. R. Webb, S. R. Payne, J. E. Wickham (29 March 1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". Br Med J (Clin Res Ed) 292 (6524): 879–882. doi:10.1136/bmj.292.6524.879. PMC 1339981. PMID 3083922. cf. Wikipedia "Simpson's Paradox"

Random Sampling

Sampling

Sampling with replacement:

1. Draw one element *uniformly at random* from list.
2. Return the element to the list.
3. Repeat

Sampling without replacement:

skip 2

What does *uniformly at random* mean? each element is equally likely

Sampling

Sampling with or without replacement:

- All samples are equally likely.
- Uniform distribution! easy to calculate $\Rightarrow$ counting

$$P(\text{sample having a certain property}) =$$

Sampling

Sampling with or without replacement:

- All samples are equally likely.
- Uniform distribution!

$$P(\text{sample having a certain property}) = \frac{\#\text{ samples having property}}{\#\text{ possible samples}}$$

Practice Problems

Example 5. There are 20 students in a class. A computer program selects a random sample of students by drawing 5 students at random **with replacement**. What is the chance that a particular student is among the 5 selected students?

Sequences of length 5

entries are in $\{1, 2, 3, \ldots, 20\}$

particular student: 17

Examples: 3, 12, 4, 9, 20

3, 3, 3, 7, 8

9, 7, 3, 3, 3

$$p = \frac{\#\text{ sequences of length 5 that include 17}}{\#\text{ sequences of length 5}}$$

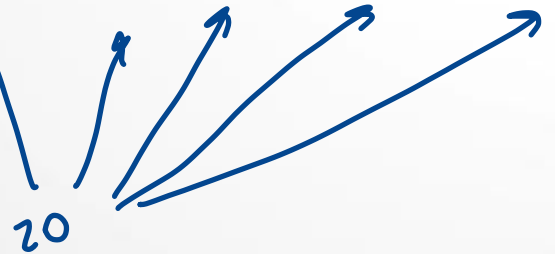
Practice Problems

Part 1. Denominator. If you draw a sample of size 5 at random with replacement from a population of size 20, how many different sequences of individuals could you draw?

sequences of length 5 with entries 1-20

$|S| = 20^5$

ex: 5 17 1 4 19



from Theory Meets Data by Ani Adhikari, Chapter 4

Practice Problems

Part 2. Numerator. If you draw a sample of size 5 at random with replacement from a population of size 20, how many different sequences of individuals include a particular person? (17)

sequences of length 5 that include 17

17 3 20 7 1

1 2 3 17 17

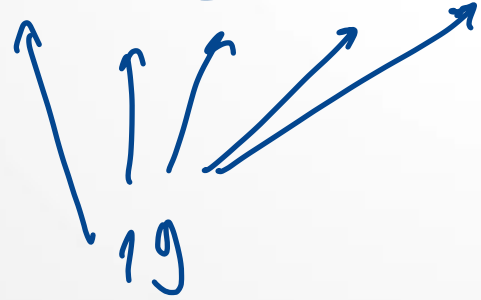
17 17 17 17 17

Practice Problems

Using the complement. If you draw a sample of size 5 at random with replacement from a population of size 20, how many different sequences of individuals **do not** include a particular person?

sequences of length 5 that don't include 17

ex: 16 2 1 4 4 $\Rightarrow 19^5$



Practice Problems

Example 5. There are 20 students in a class. A computer program selects a random sample of students by drawing 5 students at random **with replacement**. What is the chance that a particular student is among the 5 selected students?

$$P(\text{sequence of length 5 including 17}) = \frac{\#\text{seq incl. 17}}{\#\text{seq in S}} = 1 - \frac{\#\text{seq not incl. 17}}{\#\text{seq in S}} =$$

$$1 - \frac{19^5}{20^5} = 1 - \left(\frac{19}{20}\right)^5 \approx 0.226 = 22.6\%$$

complement rule

$$P(A) = 1 - P(\bar{A})$$

from Theory Meets Data by Ani Adhikari, Chapter 4

Practice Problems

Example 6. There are 20 students in a class. A computer program selects a random sample of students by drawing 5 students at random **without replacement**. What is the chance that a particular student is among the 5 selected students?

(17)

ex ✓
16, 17, 18, 19, 20

ex ✗
17, 17, 17, 17, 17

Which probability will be higher?

A. Probability of including a particular student when sampling with replacement.

B. Probability of including a particular student when sampling without replacement.

C. Both probabilities are the same.

S: sequences of length 5 without repeats

Practice Problems

Part 1. Denominator. If you draw a sample of size 5 at random without replacement from a population of size 20, how many different sequences of individuals could you draw?

sequences in S (length 5 from 1-20 without replacement)

ex: 6 17 12 9 1

20 19 18 17 16

$20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 = \frac{20!}{15!}$

$= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots \cdot 15 \cdot 16 \; 17 \cdot 18 \cdot 19 \cdot 20}{1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots \cdot 14 \cdot 15}$

from Theory Meets Data by Ani Adhikari, Chapter 4

Practice Problems

Part 2. Numerator. If you draw a sample of size 5 at random without replacement from a population of size 20, how many different sequences of individuals include a particular person?

ex: 2, 3, 4, 5, 17

3, 4, 5, 17, 2

→ 17 _ _ _ _ $19 \cdot 18 \cdot 17 \cdot 16$

→ _ 17 _ _ _ $19 \cdot 18 \cdot 17 \cdot 16$

_ _ 17 _ _ → — " —

_ _ _ 17 _ → — " —

_ _ _ _ 17 → — " —

$5 \cdot 19 \cdot 17 \cdot 17 \cdot 16 = 5 \cdot \frac{19!}{15!}$

from Theory Meets Data by Ani Adhikari, Chapter 4

Practice Problems

Using the complement. If you draw a sample of size 5 at random without replacement from a population of size 20, how many different sequences of individuals **do not** include a particular person?

sequences of length 5 not including 17

ex $\underline{2}\ \underline{1}\ \underline{20}\ \underline{18}\ \underline{3}$ $\Rightarrow 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 = \frac{19!}{14!}$

19 18 17 16 15

from Theory Meets Data by Ani Adhikari, Chapter 4

Practice Problems

Example 6. There are 20 students in a class. A computer program selects a random sample of students by drawing 5 students at random **without replacement**. What is the chance that a particular student is among the 5 selected students?

$$P\left(\begin{array}{l}\text{selecting 17 in}\\ \text{a seq. of length}\\ 5 \text{ without}\\ \text{replacement}\end{array}\right) = \frac{\#\text{ seq in } S \text{ incl. } 17}{\#\text{ seq in } S} = 1 - \frac{\#\text{ seq. not incl. } 17}{\#\text{ seq in } S} =$$

$$= \frac{5 \cdot \frac{19!}{15!}}{\frac{20!}{15!}} = 5 \cdot \frac{19!}{20!} = \frac{5}{20} = 25\% > 22.6\%$$

$$\frac{(n-1)!}{n!} = \frac{1}{n}$$

$$= 1 - \frac{19 \cdot 19 \cdot 17 \cdot 16 \cdot 15}{20 \cdot 19 \cdot 19 \cdot 17 \cdot 16} =$$

$$= \frac{20 \cdot 19 \cdot 19 \cdot 17 \cdot 16 - 19 \cdot 19 \cdot 17 \cdot 16 \cdot 15}{20 \cdot 19 \cdot 19 \cdot 17 \cdot 16} =$$

$$= \frac{20 - 15}{20} = \frac{5}{20} = 25\%$$

from Theory Meets Data by Ani Adhikari, Chapter 4

Summary

- When we sample uniformly, whether with or without replacement, each possible sample is equally likely.
- Probability questions become counting questions:

$$P(\text{sample having a certain property}) = \frac{\#\text{ samples having property}}{\#\text{ possible samples}}$$

- **Next time:** combinatorics, or counting principles