

5. The Matching Problem

Definitions and Notation

The Matching Experiment

The **matching experiment** is a **random experiment** that can be formulated in a number of colorful ways:

- Suppose that n married couples are at a party and that the men and women are randomly paired for a dance. A match occurs if a married couple happens to be paired together.
- An absent-minded secretary prepares n letters and envelopes to send to n different people, but then randomly stuffs the letters into the envelopes. A match occurs if a letter is inserted in the proper envelope.
- n people with hats have had a bit too much to drink at a party. As they leave the party, each person randomly grabs a hat. A match occurs if a person gets his or her own hat.

These experiments are clearly equivalent from a mathematical point of view, and correspond to selecting a random permutation $X = (X_1, X_2, \dots, X_n)$ of the population $D_n = \{1, 2, \dots, n\}$:

Here is the interpretation for the examples above:

- Number the couples from 1 to n . Then X_i is the number of the woman paired with the i^{th} man.
- Number the letters and corresponding envelopes from 1 to n . Then X_i is the number of the envelope containing the i^{th} letter.
- Number the people and their corresponding hats from 1 to n . Then X_i is the number of the hat chosen by the i^{th} person.

Our modeling assumption, of course, is that X is uniformly distributed on the sample space of permutations of D_n . The number of objects n is the basic parameter of the experiment. We will also consider the case of sampling *with* replacement from the population D_n , because the analysis is much easier but still provides insight. In this case, X is a sequence of independent random variables, each uniformly distributed over D_n .

Matches

We will say that a **match** occurs at position j if $X_j = j$. Thus, number of matches is the random variable N_n defined mathematically by

$$N_n = \sum_{j=1}^n I_j$$

where $I_j = \mathbf{1}(X_j = j)$ is the **indicator variable** for the event of match at position j . Our problem is to

compute the probability distribution of the number of matches. This is an old and famous problem in probability that was first considered by **Pierre-Remond Montmort**; it sometimes referred to as **Montmort's matching problem** in his honor.

Sampling With Replacement

First let's solve the matching problem in the easy case, when the sampling is with replacement.

1. Show that the sample (I_1, I_2, \dots, I_n) is a sequence of n **Bernoulli trials**, with success probability $\frac{1}{n}$.

2. Conclude that the number of matches N_n has a **binomial distribution** with trial parameter n and success parameter $\frac{1}{n}$.

$$\mathbb{P}(N_n = k) = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k}, \quad k \in \{0, 1, \dots, n\}$$

3. Use a basic results for the binomial distribution to show that

- a. $\mathbb{E}(N_n) = 1$
- b. $\text{var}(N_n) = 1$

Thus, the expected value and the variance of the number of matches are both 1, regardless of n , a somewhat surprising result at first.

4. Use a basic limit from calculus to show that $\mathbb{P}(N_n = k) \rightarrow \frac{e^{-1}}{k!}$ as $n \rightarrow \infty$ for $k \in \mathbb{N}$,

As a function of k , the right-hand side of this expression is the probability density function of the **Poisson distribution** with parameter 1. Thus, the distribution of the number of matches converges to the Poisson distribution with parameter 1 as the n increases. This is a special case of the **convergence of the binomial distribution to the Poisson**.

Sampling Without Replacement

Now let's consider the case of real interest, when the sampling is without replacement, so that X is a random permutation of the elements of D_n .

Counting Permutations with Matches

To find the **probability density function** of N_n , we need to count the number of **permutations** of D_n with a specified number of matches. This will turn out to be easy once we have counted the number of permutations with no matches; these are called **derangements** of D_n . We will denote the number of permutations of D_n with exactly k matches by

$$b_n(k) = \#(\{N_n = k\}), \quad k \in \{0, 1, \dots, n\}$$

Thus, $b_n(0)$ is the number of derrangements of D_n .

5. Show that the number of derrangements is

$$b_n(0) = n! \sum_{j=0}^n \frac{(-1)^j}{j!}$$

- Use basic properties of [counting measure](#) to show that $b_n(0) = n! - \#(\bigcup_{i=1}^n \{X_i = i\})$
- Use the [inclusion-exclusion formula](#) of combinatorics to show that

$$b_n(0) = n! - \sum_{j=1}^n (-1)^{j-1} \sum_{\#(J)=j} \#(\{X_i = i \text{ for all } i \in J\})$$
- Show that if $J \subseteq D_n$ with $\#(J) = j$ then $\#(\{X_i = i \text{ for all } i \in J\}) = (n-j)!$
- Recall that the number of subsets J of D_n with $\#(J) = j$ is $\binom{n}{j}$

6. Use the [multiplication principle](#) of combinatorics to show that

$$b_n(k) = \frac{n!}{k!} \sum_{j=0}^{n-k} \frac{(-1)^j}{j!}, \quad k \in \{0, 1, \dots, n\}$$

- A two-step procedure generates all permutations with exactly k matches.
- Select the k integers that will match. The number of ways of performing this step is $\binom{n}{k}$
- Select a permutation of the remaining $n-k$ integers with no matches. The number of ways of performing this step is $b_{n-k}(0)$

The Probability Density Function

7. Use the result of the previous exercise to show that N_n has the probability density function given below. Compare with the result in [Exercise 2](#), when the sampling is [with replacement](#).

$$\mathbb{P}(N_n = k) = \frac{1}{k!} \sum_{j=0}^{n-k} \frac{(-1)^j}{j!}, \quad k \in \{0, 1, \dots, n\}$$

8. In the [matching experiment](#), vary the parameter n and note the shape and location of the probability density function. For selected values of n , run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of empirical density function to the true probability density function.

9. Show that $\mathbb{P}(N_n = n-1) = 0$. Give a *probabilistic* proof by arguing that the event is impossible and an *algebraic* proof using the probability density function in [Exercise 7](#).

10. Use a standard infinite series from calculus to show that $\mathbb{P}(N_n = k) \rightarrow \frac{e^{-1}}{k!}$ as $n \rightarrow \infty$ for $k \in \mathbb{N}$.

Just as in the case of [sampling with replacement](#), the distribution of the number of matches converges to the [Poisson distribution](#) with parameter 1 as the n increases. The convergence is remarkably rapid. Indeed, the distribution of the number of matches with $n = 10$ is essentially the same as the distribution of the number of

matches with $n = 1000000$!

11. In the **matching experiment**, increase n and note how the probability density function stabilizes rapidly. For selected values of n , run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of the relative frequency function to the probability density function.

Moments

The **mean** and **variance** of the number of matches could be computed directly from the distribution. However, it is much better to use the representation in terms of **indicator variables**. The **exchangeable property** is an important tool in this section.

12. Show that $\mathbb{E}(I_j) = \frac{1}{n}$ for any $j \in \{1, 2, \dots, n\}$.

13. Show that $\mathbb{E}(N_n) = 1$. *Hint:* Use the result of Exercise 12 and basic **properties of expected value**.

Thus, the expected number of matches is 1, regardless of n , just as when the sampling is **with replacement**.

14. Show that $\text{var}(I_j) = \frac{n-1}{n^2}$ for any $j \in \{1, 2, \dots, n\}$.

A match in one position would seem to make it more likely that there would be a match in another position. Thus, we might guess that the indicator variables are positively **correlated**.

15. Show that for distinct j and k in $\{1, 2, \dots, n\}$,

a. $\text{cov}(I_j, I_k) = \frac{1}{n^2(n-1)}$

b. $\text{cor}(I_j, I_k) = \frac{1}{(n-1)^2}$

From Exercise 15, when $n = 2$, the event that there is a match in position 1 is perfectly correlated with the event that there is a match in position 2. Does this result seem reasonable?

16. Show that $\text{var}(N_n) = 1$. *Hint:* Use the previous two exercises and basic properties of covariance.

17. In the **matching experiment**, vary the parameter n and note the shape and location of the mean/standard deviation bar. For selected values of the parameter, run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of sample mean and standard deviation to the distribution mean and standard deviation.

18. Show that for distinct j and k in $\{1, 2, \dots, n\}$, $\text{cov}(I_j, I_k) \rightarrow 0$ as $n \rightarrow \infty$.

Thus, the event that a match occurs in position j is nearly independent of the event that a match occurs in position k if n is large. For large n , the indicator variables behave nearly like n **Bernoulli trials** with success probability $\frac{1}{n}$, which of course, is what happens when the sampling is **with replacement**.

A Recursion Relation

In this subsection, we will give an alternate derivation of the distribution of the number of matches, in a sense by **embedding** the experiment with parameter n into the experiment with parameter $n + 1$. First, consider the random permutation $(X_1, X_2, \dots, X_n, X_{n+1})$ of D_{n+1}

19. Argue that (X_1, X_2, \dots, X_n) is a random permutation of D_n if and only if $X_{n+1} = n + 1$ if and only if $I_{n+1} = 1$

20. Use the result of the previous exercise to argue that

$$\mathbb{P}(N_n = k) = \mathbb{P}(N_{n+1} = k + 1 | I_{n+1} = 1), \quad k \in \{0, 1, \dots, n\}$$

21. Use the result of the previous exercise and a conditional probability argument to show that

$$\mathbb{P}(N_n = k) = \mathbb{P}(N_{n+1} = k + 1) \frac{\mathbb{P}(I_{n+1} = 1 | N_{n+1} = k + 1)}{\mathbb{P}(I_{n+1} = 1)}, \quad k \in \{0, 1, \dots, n\}$$

22. Argue that

a. $\mathbb{P}(I_{n+1} = 1) = \frac{1}{n+1}$

b. $\mathbb{P}(I_{n+1} = 1 | N_{n+1} = k + 1) = \frac{k+1}{n+1}$

23. Finally conclude that $\mathbb{P}(N_n = k) = (k + 1) \mathbb{P}(N_{n+1} = k + 1)$ for $k \in \{0, 1, \dots, n\}$

24. Note also that $\mathbb{P}(N_1 = 1) = 1$.

The results of the previous two exercises can be used to obtain the probability density function of N_n recursively for any n .

The Probability Generating Function

Next recall that the **probability generating function** of N_n is given by

$$G_n(t) = \mathbb{E}(t^{N_n}) = \sum_{j=0}^n \mathbb{P}(N_n = j) t^j, \quad t \in \mathbb{R}$$

25. Use the results of the last subsection and basic calculus to show that the family of probability generating functions satisfies the following differential equations and ancillary conditions:

$$\left(\frac{dG_{n+1}}{dt} \right) (t) = G_n(t), \quad G_{n+1}(1) = 1$$

Note that $G_1(t) = t$ for $t \in \mathbb{R}$. This fact, together with the system of differential equations in the previous exercise, can be used to compute G_n for any n

26. Show that for $t \in \mathbb{R}$,

- a. $G_2(t) = \frac{1}{2} + \frac{1}{2}t^2$
- b. $G_3(t) = \frac{1}{3} + \frac{1}{2}t + \frac{1}{6}t^3$
- c. $G_4(t) = \frac{3}{8} + \frac{1}{3}t + \frac{1}{4}t^2 + \frac{1}{24}t^4$

27. Use [Exercise 25](#) to show that

$$\frac{d^k}{dt^k} G_n(t) = G_{n-k}(t)$$

28. Use the result of the previous exercise and basic properties of generating functions to conclude that

$$\mathbb{P}(N_n = k) = \frac{1}{k!} \mathbb{P}(N_{n-k} = 0), \quad k \in \{0, 1, \dots, n\}$$

Examples and Applications

29. A secretary randomly stuffs 5 letters into 5 envelopes.

- a. Compute the number of outcomes with exactly k letters in the correct envelope, for each $k \in \{0, 1, 2, 3, 4, 5\}$
- b. Find the probability density function of the number of letters in the correct envelopes.
- c. Compute the covariance and correlation of the events that a letter is in the correct envelope and another letter is in the correct envelope.



30. Ten married couples are randomly paired for a dance.

- a. Find the probability density function of the number of married couples who are paired together.
- b. Find the mean and variance of the number of married couples who are paired together.
- c. Find the probability that at least 3 couples are paired together.



31. In the [matching experiment](#), set $n = 10$. Run the experiment 1000 times with an update frequency of 10. Compare the following for the number of matches:

- a. the true probabilities
- b. the relative frequencies from the simulation
- c. the limiting Poisson probabilities

