Random Sampling
(Drawing a Random Sample from a Population)

If we toss a 40¢ piece, there is no real "population" that we're "sampling from." There is only one such coin on Earth, and it might never be tossed again. The "probability of heads," θ, represents its tendency to fall heads.

The students in this class, on the other hand, do represent a population (of size, say, N = 200). There are many ways that I can select a sample from this population. For a sample of size one, I could

... ask for a volunteer
... pick the closest student
... pick the one who overslept and missed the midterm exam
... choose one "at random"

By "choose one student at random" we mean selecting one by means of a random process, such as:

- Write each name on a piece of paper, put them all into a hat, stir thoroughly, and draw one.

- Number the students 1,2,...,200. Put 200 ping pong balls numbered 1-200 into a large basket, mix thoroughly, and draw one. Select the student whose number is drawn.

Now, if N_D is the number who are in doctoral programs, then the probability of selecting a doctoral student from the class, θ, equals $N_D/200$, the proportion of doctoral students in the class.

We are performing a Bernoulli trial when we use one of the above procedures to select a single student and then observe whether he/she is in a doctoral degree program or not.

⇨ By the process of selecting at random, we <u>make</u> the <u>proportion</u>, $N_D/200$, into a <u>probability</u>.

- The <u>proportion</u> of doctoral students in the class is $N_D/200$.

- The <u>probability</u> that a student in this class is a doctoral student is unknown!

- The <u>probability</u> that a randomly selected student from this class is a doctoral student is just the proportion of doctoral students, $N_D/200$.

Example
Consider the following students:

- the student who overslept and missed the midterm exam
- the student whose name appears first on the class list
- the student who volunteered

It might make sense to speak of the probability that one of these students is a doctoral student,
but it is not at all clear how that probability would ever be determined, or that is has any objective, scientifically useful meaning.

If the proportion of doctoral students is 1/5, does that make the probability 1/5 that the student who overslept is a doctoral student?

Maybe doctoral students are obsessive about test-taking, while those in masters degree programs are happy-go-lucky slackers, for the most part. If this is true, then the probability that the student selected by this process (pick the one who overslept) is a doctoral student might be much less than 1/5.

Suppose we randomly select a student using the Ping pong ball routine. Now, each student in the class has the same chance of being randomly selecting, 1/200.

Now we **independently** repeat this selection process 10 times. (This means that any student can be selected multiple times).

Then the number of selected students who are doctoral candidates has a Binomial distribution.

That is,

Y = number of doctoral students selected
$Y \sim bin(10, N_D/200)$

It is absolutely crucial to understand that independently repeating this experiment entails using all 200 ping pong balls on each trial. We do not exclude those students previously selected.

⇨ This type of sampling is called **simple random sampling with replacement**.

We can also use simple random sampling **without replacement**.

If we don't replace the ping pond ball (or slip or paper) after each draw, then the Binomial model is not longer the correct model for Y (number of doctoral students selected).

Here is why:

If X_1 and X_2 represent the results of the two draws, then

$$P(X_2 = 1 \mid X_1 = 1) = (N_D-1)/199$$

$$P(X_2 = 1 \mid X_1 = 0) = N_D/199$$

We see that $P(X_2 = 1 \mid X_1)$ depends on the value of X_1. If $X_1 = 1$, it has one value, but if $X_1 = 0$, it has another.

Therefore X_1 and X_2 are not independent. Because the trials are not independent, the Binomial model does not apply here.

In fact, we have two Bernoulli trials, both with probability of success $N_D/200$.

How can this be?

Clearly the first trial has success probability $N_D/200$.

But how about the second one?

The probability of success on the second trial, i.e. selecting a doctoral student on the second draw is

$$\begin{aligned} P(X_2=1) &= P(X_2=1 \mid X_1=0)P(X_1=0) \\ &\quad + P(X_2=1 \mid X_1=1)P(X_1=1) \\ &= (N_D/199)(1-N_D/200) \\ &\quad + ((N_D-1)/199)(N_D/200) \\ &= N_D/200 \end{aligned}$$

Each draw is a Bernoulli trial and they both have the same success probability: $P(X_1=1) = P(X_2=1) = N_D/200$, but they are not independent trials. What happens on the first influences the chance of success on the second.

⇨ This independence thing is important!

When we speak of random sampling from a real population like this, and we are considering more than one draw, we must be careful to distinguish between sampling with replacement and sampling without replacement.

"With replacement" makes the results of each draw independent of the others.

"Without replacement" makes the results of different draws dependent.

Simple Random Sampling Without Replacement

The distinction between sampling from an actual population (like this class, or the people of London) and processes like tossing a coin, or observing the sex of a baby, is an important one.

In sampling from an actual population it is important whether the sampling is with or without replacement. But in coin tossing or observing the sex of a baby, it doesn't even make sense to ask whether the sampling was with replacement.

⇨ The "randomness" in a sequence of coin tosses is part of the very nature of the process. In fact it is unavoidable. But if you want a random sample from a real population of people or things, you must make it random by the way you choose which ones to include in the sample, and which ones will be left out.

Note: It is common to refer to real populations, like the students in this class, as "finite" populations. Processes like observing tosses of a coin are then described as "drawing a random sample from an infinite population".

We saw that the Binomial distribution can be used to model the result of SRS with replacement.

(1) The "SRS" is important because it implies that each trial has the same probability of success, in our case $N_D/200$.

(2) The "with replacement" is important because it implies that each trial is independent.

Numbers (1) and (2) **together** imply the binomial distribution is the correct model for our random variable, Y, representing the number of doctoral students selected in our sample.

It is possible to change (1) so that everyone has some fixed probability of being selected, but the probabilities are not the same for everyone.

For example, I could increase the selection probability of some people by putting extra ping pong balls in the cage.

These are still random sampling procedures, because everyone has a fixed probability of being selected. We won't deal with it much more in this class, just remember that the probabilities do not all have to be equal.

Suppose we use SRS **without** replacement to select a sample of size n from a population of N objects.

There are $\binom{N}{n}$ possible samples, and each sample has the same probability, $1/\binom{N}{n}$, of being selected.

Example
Draw a simple random sample of size 20 (without replacement) from this class of 200 students. What is your probability of being in the selected sample?

Every sample has probability $1/\binom{N}{n} = 1/\binom{200}{20}$.

P(You're in) = {number of samples you're in}/ $\binom{200}{20}$

Every sample that you are in consists of you and 19 other students. The number of those samples is just the number of ways we can select the other 19 to be in the sample with you, out of the other 199 students left in the class,

{number of samples you're in} = $\binom{199}{19}$.

So,

$$\text{P(You're in)} = \binom{199}{19} / \binom{200}{20} = \frac{199!}{19!\,180!} \frac{20!\,180!}{200!}$$

$$= 20/200$$

$$= 1/10.$$

Example (continued)
What is the answer to the previous question if I sample with replacement?

P(You're in) = 1- P(You're out on all 20 draws)

(because you will be in the selected sample unless you're out on every draw)

Now the probability that you're out on one draw is

$$P(\text{Out on one draw}) = (N-1)/N$$

Since we're sampling with replacement, the draws are independent, so

P(You're out on all 20 draws)

$$= P(\text{Out on 1st and Out on 2nd ... and Out on 20th})$$

$$= ((N-1)/N)^{20}$$

$$P(\text{You're in}) = 1 - ((N-1)/N)^{20}$$

$$= 1 - (199/200)^{20}$$

Without replacement: P(You're in) = 0.100
With replacement: P(You're in) = 0.0954

Let's try a harder problem

Draw a simple random sample (without replacement) of size n = 5 from this class (population) of N = 200 students. What is the probability that I will find no doctoral students?

(a) From SRS, every possible sample has the same probability of being selected. There are

$$\binom{200}{5} = 2{,}535{,}650{,}040$$

possible samples, so each sample has probability 1 / 2,535,650,040 of being selected.

(b) There are (200-N_D) non-doctoral students in the class, and therefore

$$\binom{200 - N_D}{5}$$

possible samples of five non-doctoral students.

(a) and (b) together imply that the probability of selecting a sample containing no doctoral students is

$$\text{P(no doctoral in sample of five)} = \binom{200 - N_D}{5} \Big/ \binom{200}{5}$$

Just to make it concrete, let's suppose that there are 35 doctoral students: $N_D = 35$. In that case, the number of samples containing no doctoral students is the number of ways to choose 5 from the (200-35) = 165 non-doctoral students, or

$$\binom{165}{5} = 958,683,033 .$$

Therefore the probability that I will select a sample containing no doctoral students is

$$\binom{165}{5} / \binom{200}{5} = \frac{958,683,033}{2,535,650,040} = 0.378 .$$

From this we get the probability that a simple random sample (without replacement) of size 5 from this class will include at least one doctoral student is

P(at least one)=1-P(none)=1-0.378 = 0.622

Q: What is the probability that the sample will contain <u>exactly</u> one doctoral student?

To answer this question we need to find how many of the 2,535,650,040 equally probable samples contain exactly one doctoral student.

There are 35 ways to pick the one doctoral student to be in the sample, and for each of those, there are $\binom{165}{4}$ ways to pick the other four sample members from among the 165 non-doctoral students.

Thus, there are

$$35\binom{165}{4} = 35(29{,}772{,}765) = 1{,}042{,}046{,}775$$

samples that contain exactly one doctoral student, and the probability of selecting one of these samples is:

$$\frac{1{,}042{,}046{,}775}{2{,}535{,}650{,}040} = 0.411 \text{ .}$$

Q: What is the probability that the sample will contain <u>exactly</u> two doctoral students?

There are $\binom{35}{2} = \frac{35!}{2!\ 32!} = \frac{35 \times 34}{2 \times 1} = 595$ ways to pick

the two doctoral students and for each of these 595 two-doctoral-student pairs, there are

$$\binom{165}{3} = \frac{165!}{3!\ 162!} = \frac{165 \times 164 \times 163}{3 \times 2 \times 1} = 735{,}130$$

ways to pick the three non-doctoral students, so the probability of getting two doctoral students in a SRS of $n = 5$ is

$$\frac{\binom{35}{2}\binom{165}{3}}{\binom{200}{5}} = \frac{(595)(735{,}130)}{2{,}535{,}650{,}040} = \frac{437{,}402{,}350}{2{,}535{,}650{,}040} = 0.173$$

Similarly, the probability of three doctorals is

$$\frac{\binom{35}{3}\binom{165}{2}}{\binom{200}{5}} = \frac{(6545)(13530)}{2{,}535{,}650{,}040} = \frac{88{,}553{,}850}{2{,}535{,}650{,}040} = 0.035,$$

The probability of four is

$$\frac{\binom{35}{4}\binom{165}{1}}{\binom{200}{5}} = \frac{(52360)(165)}{2{,}535{,}650{,}040} = \frac{8{,}639{,}400}{2{,}535{,}650{,}040} = 0.003,$$

and the probability of five is

$$\frac{\binom{35}{5}\binom{165}{0}}{\binom{200}{5}} = \frac{(324632)(1)}{2{,}535{,}650{,}040} = 0.00013$$

If *X* represents the number of doctoral students in a SRS (without replacement) from this class, we have found that

$$
\begin{aligned}
P(X = 0) &= 0.378 \\
P(X = 1) &= 0.411 \\
P(X = 2) &= 0.173 \\
P(X = 3) &= 0.035 \\
P(X = 4) &= 0.003 \\
P(X = 5) &= 0.000 \\
&\quad 1.000
\end{aligned}
$$

These probabilities are given by the formula:

$$P(X = k) = \frac{\binom{35}{k}\binom{200-35}{5-k}}{\binom{200}{5}} \qquad \text{for } k = 0, 1, 2, 3, 4, 5$$

This probability distribution also has a name; it is a **Hypergeometric** probability distribution.

The Hypergeometric Probability Distribution

For a population of N objects, of which, we sample n of them without replacement and C of the objects have some characteristic, then

(1) X = the number of sample observations having that characteristic

(2) S = {max(0,n-(N-c)), …, min(n,C)}

(3) For some k: {max(0,n-(N-C) ≤ k ≤ min(n,c)}

$$P(X = k) = \frac{\binom{C}{k}\binom{N-C}{n-k}}{\binom{N}{n}}$$

X cannot be bigger than n nor can it be bigger than *C*. Therefore the upper limit of S is min(n,c).

For the lower limit: (*N*-*C*) is the number in the population without the characteristic, and if the sample size exceeds this number by one, then there must be at least one unit with the characteristic in the sample. If the sample size exceeds this number by two, then there must be at least two units with the characteristic in the sample. Etc…

Example
I have a jar containing 10 marbles, of which 7 are black and 3 are white. If I draw a SRS (without replacement) of 2 marbles, what is the probability distribution of the number of black marbles in the sample?

- The number of black marbles in the sample is a random variable, X, with a Hypergeometric probability distribution.

$$P(X = k) = \frac{\binom{7}{k}\binom{10-7}{2-k}}{\binom{10}{2}} \qquad \text{for } k = 0,1,2$$

For instance, $P(X = 0) = \frac{\binom{7}{0}\binom{3}{2}}{\binom{10}{2}} = \frac{3}{45} = 0.067.$

ASIDE:
If I replace the marble after each draw (i.e. sample with replacement) the number of black marbles that I will see in two draws, say Y, has a binomial probability distribution with n = 2 and θ = 0.7, so

$$P(Y = 0) = \binom{2}{0}(0.7)^0(1-0.7)^2 = (0.3)^2 = 0.090.$$

Consider n = 5 under the without replacement sampling scheme.

Now X cannot be zero -- there are only three white marbles, so if I draw five I *must* get at least two black ones; therefore k = 0 and k = 1 are impossible.

Here N = 10, n = 5, and C = 7, and the smallest possible value for X is: n- (N- C) = 5-(10-7) = 2.
The largest possible value is min(n,C)=min(5, 7)=5.

- The probability distribution is given by

$$P(X = k) = \frac{\binom{7}{k}\binom{10-7}{5-k}}{\binom{10}{5}} \qquad \text{for } k = 2, 3, 4, \text{ and } 5.$$

Aside:
If I sample with replacement, then the probability of one black ball in 5 draws is

$$P(Y = 1) = \binom{5}{1}(0.7)^1(1-0.7)^4 = 0.028 ,$$

but for without replacement sampling, this result is impossible: P(X = 1) = 0.