

1 Econ 7818 a few questions on populations, sampling, joint density function of the sample, etc.

Set 4, Dec 27, 2009

1. Assume a rv X with some population density function $f_X(x)$. Assume a random sample of size n from this density function. Part 1: Specify and describe the joint density function for random samples of size n taken from this population. A statistic of a sample of size n is of the form $s = s(X_1, X_2, \dots, X_n)$, where X_j is the j^{th} draw. Describe to the reader how you would, in general terms, determine the expected value of s , $E[s]$, and the variance of s , σ_s^2

Part 2: Now choose and specify a functional form for $f_X(x)$; choose a density function with one or two parameters. Do not choose the normal. Specify the joint density function of samples of size n given your specific $f_X(x)$.

Part 3: Specify a specific functional form for $s = s(X_1, X_2, \dots, X_n)$ - not the sample mean. Determine $E[S]$ and σ_S^2 ; if you can't determine $E[S]$ and σ_S^2 , at least show the math you would need to solve.

Part 4: Now assume specific numerical values for the parameter in your $f_X(x)$, assume $n = 25, 50$ or 100 , and pull 100 random samples. (Don't force your reader to look at all of these numbers.) Briefly describe the process you used to derive your random samples.

Part 5: Calculate and plot the 100 values of your S statistic. That is, show me the simulated sampling distribution for your statistic. Does the distribution look as you expected? Choose some specific number, m , and determine the probability that $s > m$.

this was a mathematica assignment 2007

2. Assume the very simple discrete-density function: $f(x) = .2$ if $x = 1$, $f(x) = .5$ if $x = 2$, $f(x) = .3$ if $x = 3$, and zero otherwise. Graph the density function. Specify and graph the CDF. Now assume a random sample of 10 observations, X_1, X_2, \dots, X_{10} from this density function. Specify the joint density function for the sample $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10})$. Interpret $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10})$ as a probability. Be as specific as you can in your interpretation.

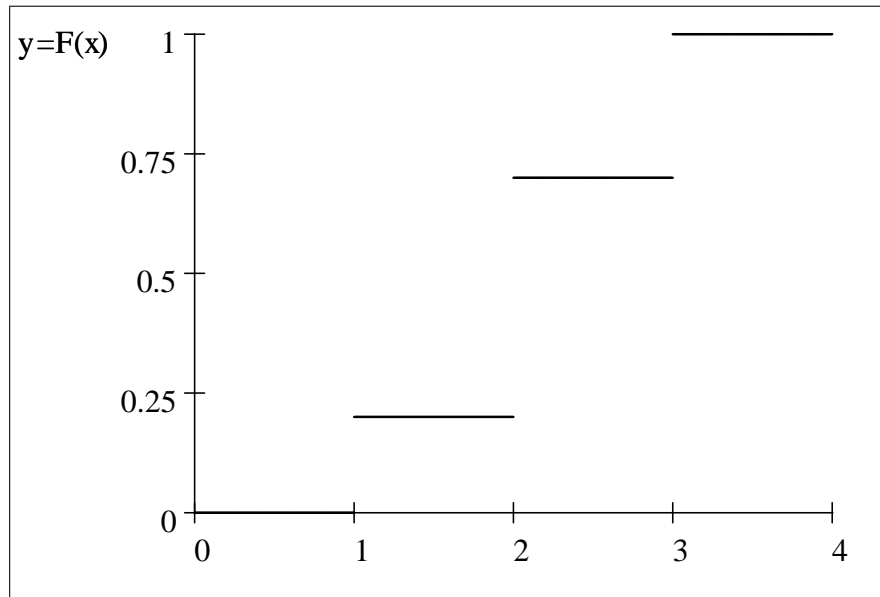
Now assume a random sample is drawn and it is 1, 1, 3, 2, 2, 2, 3, 1, 2, 1. In this sample, there are four 1's, four 2's, and two 3's. What is the probability that you would draw this sample.

What sample has the highest probability of being drawn when the sequence in which the ones, twos and threes matter. What is this sample and what

is the probability of drawing it. What sample has the highest probability of being drawn when the sequence of ones, two and threes drawn does not matter. What is the probability of this sample and what is the probability of drawing it.

a partial answer: The CDF is

$$\begin{cases} 0 & \text{if } x < 1 \\ .2 & \text{if } 1 \leq x < 2 \\ .7 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$



The joint density function for the sample is $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10}) = .2^{N_1} .5^{N_2} .3^{(10-N_1-N_2)}$ where N_1 is the number of 1's, N_2 is the number of 2's, and $(10-N_1-N_2)$ is the number of 3's observed. Note that since this is a discrete distribution $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10}) = .2^{N_1} .5^{N_2} .3^{(10-N_1-N_2)}$ is the probability of observing the sample X_1, X_2, \dots, X_{10} .¹ Note that this probability is for for the ones, twos and threes in the sample **in the order they are sampled**.

The probability of observing drawing the sample $(1, 1, 3, 2, 2, 2, 3, 1, 2, 1)$ (different sequences are different samples) is $f_{X_1, X_2, \dots, X_{10}}(1, 1, 3, 2, 2, 2, 3, 1, 2, 1) = (.2)^4 (.5)^4 (.3)^2 = 9.0 \times 10^{-6}$, not very high.²

¹If the density function is discrete, we can strictly talk about the probability of observing a sample rather than just the *likelihood* of observing a sample. If the density function is continuous, we are limited to talking only about the likelihood of a sample being drawn.

²Keep in mind the distinction between the probability of observing a particular sequence of observations and the probability of observing a sample with a given number of 1's, 2's and

Which sample(s) have the highest probability of being pulled when sequence matters? It is those that have the N_1 and N_2 that maximizes $(.2)^{N_1}(.5)^{N_1}(.3)^{1-N_1-N_2}$. There is only one, $N_2 = 10$ (all 2's are drawn). The probability of drawing this sample is $(.2)^0(.5)^{10}(.3)^0 = 9.7656 \times 10^{-4}$. You might find this surprising. Note that the probability of drawing this sample is extremely low.

Contrast the sample with the highest probability of being observed when sequence matters (all twos) with the sample with the highest probability of being observed when sequence does not matter - only the number of 1's, 2's and 3's matter. To find this probability one maximizes $\frac{10!}{N_1!N_2!(10-N_1-N_2)!}(.2)^{N_1}(.5)^{N_1}(.3)^{1-N_1-N_2}$ with respect to N_1 and N_2 . The probability is maximized when $N_1 = 2$ and $N_2 = 5$, and the maximum probability is $\frac{10!}{2!5!(3)!}(.2)^2(.5)^5(.3)^3 = 2520(3.375 \times 10^{-5}) = 0.08505$.

3. Assume the very simple discrete-density function: $f(x) = p_1$ if $x = 1$, $f(x) = p_2$ if $x = 2$, $f(x) = (1 - p_1 - p_2)$ if $x = 3$, and zero otherwise. Assume $1 > p_1, p_2 > 0$ such that $(1 - p_1 - p_2) > 0$. Now assume a random sample of 10 observations, X_1, X_2, \dots, X_{10} from this density function. Specify the joint density function for the sample $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10})$. Now assume that a random sample is drawn and it is 3, 2, 3, 2, 2, 2, 3, 1, 2, 1. In this sample, there are two 1's, five 2's, and three 3's. Find those values of α and B that maximize the probability of drawing this particular sample. Then find those values of α and B that maximize the probability of getting drawing a sample with four 1's, four 2's, and two 3's.

The joint density of the the sample is $f_{X_1, X_2, \dots, X_{10}}(X_1, X_2, \dots, X_{10}) = (p_1)^{N_1}(p_2)^{N_2}(1 - p_1 - p_2)^{10 - N_1 - N_2}$ where where N_1 is the number of 1's, N_2 is the number of 2's, and $(10 - N_1 - N_2)$ is the number of 3's observed. Find the p_1 and p_2 that maximize $(p_1)^3(p_2)^5(1 - p_1 - p_2)^2$. The maximizing values are .3, .5 and .2. This can be seen with a grid search over values of p_1 and p_2

$$(.1)^2(.1)^5(1 - .1 - .1)^3 = 5.12 \times 10^{-8}$$

$$(.1)^2(.2)^5(1 - .1 - .2)^3 = 1.0976 \times 10^{-6}$$

$$(.1)^2(.3)^5(1 - .1 - .3)^3 = 5.2488 \times 10^{-6}$$

$$(.1)^2(.4)^5(1 - .1 - .4)^3 = 1.28 \times 10^{-5}$$

$$(.1)^2(.5)^5(1 - .1 - .5)^3 = 0.00002$$

3's - often many different sequences can generate the same number of 1's, 2's and 3's.

Note that there are $\frac{10!}{4!4!2!} = 3150$ different sequences that generate 4 1's, 4 2's and 2 2's, so while the probability associated with each specific sequence of 4 1's, 4 2's and 2 2's is 9.0×10^{-6} , the probability of observing 4 1's, 4 2's and 2 2's is $3150(9.0 \times 10^{-6}) = 0.02835$.

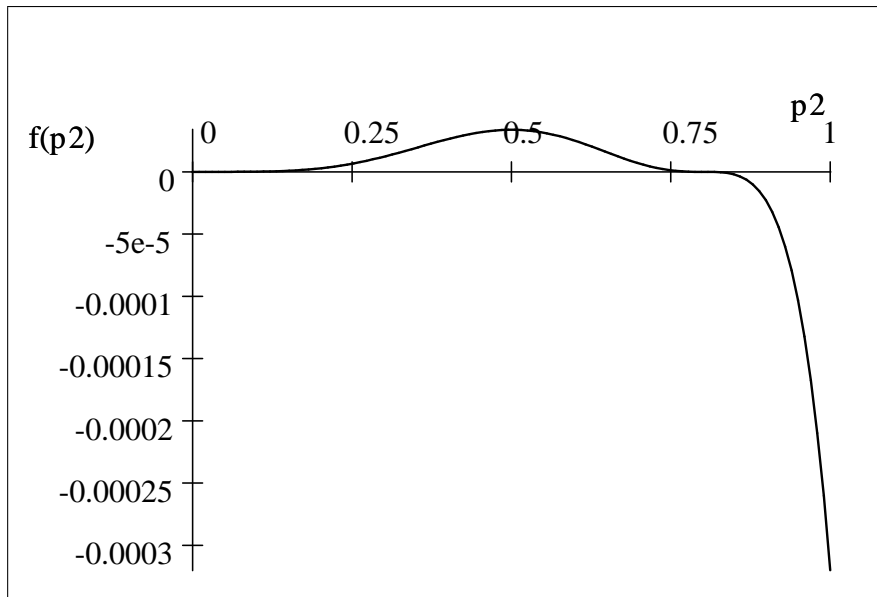
So, when we talk about the probability of observing a sample we need to distinguish between the probability of observing a sample when the sequence matters and the probability of observing a sample if the sequence does not matter.

$$\begin{aligned}
(.1)^2(.6)^5(1 - .1 - .6)^3 &= 2.0995 \times 10^{-5} \\
(.1)^2(.7)^5(1 - .1 - .7)^3 &= 1.3446 \times 10^{-5} \\
(.1)^2(.8)^5(1 - .1 - .8)^3 &= 3.2768 \times 10^{-6} \\
(.2)^2(.1)^5(1 - .2 - .1)^3 &= 1.372 \times 10^{-7} \\
(.2)^2(.2)^5(1 - .2 - .2)^3 &= 2.7648 \times 10^{-6} \\
(.2)^2(.3)^5(1 - .2 - .3)^3 &= 1.215 \times 10^{-5} \\
(.2)^2(.4)^5(1 - .2 - .4)^3 &= 2.6214 \times 10^{-5} \\
(.2)^2(.5)^5(1 - .2 - .5)^3 &= 3.375 \times 10^{-5} \text{ the largest one.} \\
(.2)^2(.6)^5(1 - .2 - .6)^3 &= 2.4883 \times 10^{-5} \\
(.2)^2(.7)^5(1 - .2 - .7)^3 &= 6.7228 \times 10^{-6} \\
(.3)^2(.1)^5(1 - .3 - .1)^3 &= 1.944 \times 10^{-7} \\
(.3)^2(.2)^5(1 - .3 - .2)^3 &= 3.6 \times 10^{-6} \\
(.3)^2(.3)^5(1 - .3 - .3)^3 &= 1.3997 \times 10^{-5} \\
(.3)^2(.4)^5(1 - .3 - .4)^3 &= 2.4883 \times 10^{-5} \\
(.3)^2(.5)^5(1 - .3 - .5)^3 &= 2.25 \times 10^{-5} \\
(.3)^2(.6)^5(1 - .3 - .6)^3 &= 6.9984 \times 10^{-6} \\
(.4)^2(.1)^5(1 - .4 - .1)^3 &= 2.0 \times 10^{-7} \\
(.4)^2(.2)^5(1 - .4 - .2)^3 &= 3.2768 \times 10^{-6} \\
(.4)^2(.3)^5(1 - .4 - .3)^3 &= 1.0498 \times 10^{-5} \\
(.4)^2(.4)^5(1 - .4 - .4)^3 &= 1.3107 \times 10^{-5} \\
(.4)^2(.5)^5(1 - .4 - .5)^3 &= 5.0 \times 10^{-6} \\
(.5)^2(.1)^5(1 - .5 - .1)^3 &= 1.6 \times 10^{-7} \\
(.5)^2(.2)^5(1 - .5 - .2)^3 &= 2.16 \times 10^{-6} \\
(.5)^2(.3)^5(1 - .5 - .3)^3 &= 4.86 \times 10^{-6} \\
(.5)^2(.4)^5(1 - .5 - .4)^3 &= 2.56 \times 10^{-6} \\
(.6)^2(.1)^5(1 - .6 - .1)^3 &= 9.72 \times 10^{-8} \\
(.6)^2(.2)^5(1 - .6 - .2)^3 &= 9.216 \times 10^{-7} \\
(.6)^2(.3)^5(1 - .6 - .3)^3 &= 8.748 \times 10^{-7} \\
(.7)^2(.1)^5(1 - .7 - .1)^3 &= 3.92 \times 10^{-8} \\
(.7)^2(.2)^5(1 - .7 - .2)^3 &: 1.568 \times 10^{-7} \\
(.8)^2(.1)^5(1 - .8 - .1)^3 &= 6.4 \times 10^{-9}
\end{aligned}$$

We know that the answer is $p_1 = .2$ and $p_2 = .5$, but let's play more.

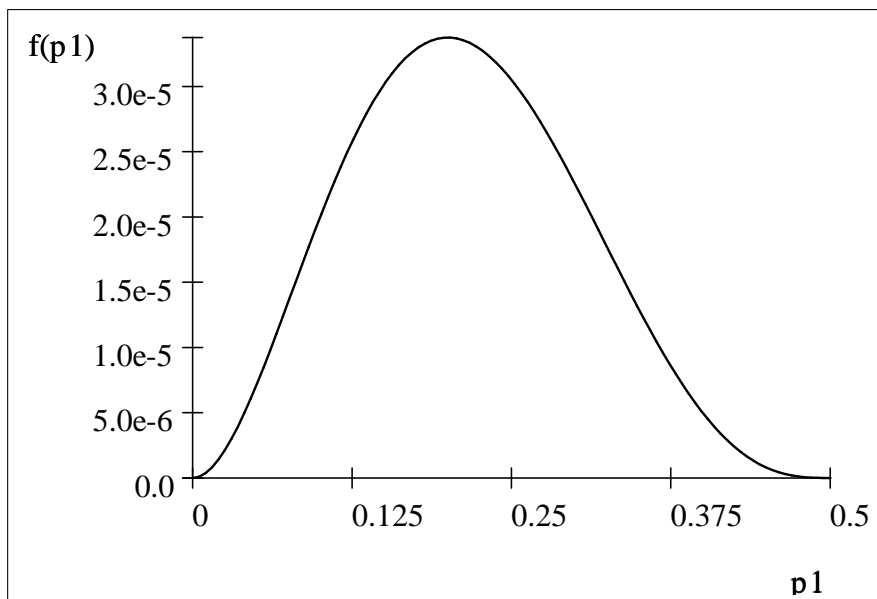
First I assumed, I knew $p_1 = .2$ and graph

$$(.2)^2(p_2)^5(1 - .2 - p_2)^3$$



The function is maximized at $p_2 = .5$ if $p_1 = .2$

Then I assumed, I knew that $p_2 = .5$ and graphed $(p_1)^2(.5)^5(1 - p_1 - .5)^3$



The function is maximized at $p_1 = .2$ if $p_2 = .5$

So, summarizing to here, $p_1 = .2$ and $p_2 = .5$ are those values that maximize the probability of observing the specific sample drawn. But, what is the probability of observing two ones, five twos and three threes, independent of the order they are drawn. It is the same: $p_1 = .2$ and $p_2 = .5$. The probability of observing two ones, five twos and three threes in any order is $\frac{10!}{2!5!(10-2-5)!} (p_1)^2 (p_2)^5 (1 - p_1 - p_2)^3$. Since p_1 and p_2 do not appear in the first term, the p_1 and p_2 that maximize $(p_1)^2 (p_2)^5 (1 - p_1 - p_2)^3$ are the same p_1 and p_2 that maximize $\frac{10!}{2!5!(10-2-5)!}$.

4. there is some other stuff I might mine from assignment 7 in 2006, see the assignment 7 tex file.

5. Give me an example of a nonrandom sample and explain why it is not random. Start by identifying the population you are sampling from. Contrast your nonrandom sample with a random sample.

answer: provide an example of a sample. Show that the process that generated the sample was sufficient to generate a random sample. Random samples can look very nonrandom. A statistics prof had his students turn in random samples, a sample of ten from the set of all integers between zero and ten. His claim was that he could tell who "cooked the books". His claimed as fakes those examples that looked random - he got it right most of the time. That said, no, one cannot tell by observation, whether a sample is a random sample? The process, not the outcome, is what determines whether a sample is a random sample. Samples are like estimates, what makes them good or bad is the process that generated them.

6. Assume a univariate random variable. Explain to the reader the density function for a random sample (*distribution of the sample*). Now specify some density function for the population that is a function of at least one parameter. Now assume you have a random sample of five observations and specify those five observations. Write down the density function for your random sample. Finally find the value of the population density parameter(s) that maximize the density function for your sample.

7. "Make up and answer a question on X that you think would be a good question to ask about X on the final. For example, write a short essay that introduces the reader to how one might derive the distribution of a statistic that is a function of one or more random variables, given knowledge of the joint density of those random variables. To answer such a question you would want to introduce the topic rather than present the most advanced stuff on the topic, and your answer would probably have a lot of explanatory words. I would want to see how you would explain in your own words.

"X", could be the above, or sampling distributions, or the central-limit theorem.

8. MGB define a random sample as: The sample (x_1, x_2, \dots, x_n) is a random sample from $f_X(x)$ if

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f_X(x_1)f_X(x_2)\dots f_X(x_n)$$

where $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ is the joint distribution of the sample (MGB 223).

With this definition in mind, discuss taking a random sample from a population with a finite number of members, m , where $m > n$.

Discussion: We need to distinguish between random sampling with and without replacement. With replacement, one can take a random sample of any size, n , (even greater than the population size) by taking n independent draws from $f_X(x)$; note that the population is unaffected by the draws, so $f_{X_1}(x_1) = f_{X_2}(x_2) = \dots = f_{X_m}(x_m) = f_X(x)$. This is not the case if one samples from a finite population without replacement.

In the words of MGB, "We might further note that our definition of random sampling has automatically ruled out sampling from finite population without replacement since, then, the results of the drawings are not independent." MGB page 225.

Put simply, it cannot be done: random sampling requires that the joint distribution of the sample be the product of the density functions for each draw, $f_{X_i}(x_i)$ for draw i , and that all of these density functions be identical, $f_{X_i}(x_i) = f_X(x) \forall i$. If one takes draws without replacement from a finite sample, one will violate these properties of required joint density function.

Give the following more thought. For example, if the population consists of 10 individuals, 5 females and 5 males and, on the first draw one draws in such a way that each member of the population has an equal chance of being drawn and a female is drawn, on the second draw one is drawing from a population with 4 females and 5 males; the probability of drawing a female has changed: $f_{X_2}(x_2) \neq f_{X_1}(x_1)$

Further thought: Some loosely state, I have, that a sample is random if each individual in a population has an equal chance of being in the sample. This is not quite correct. It is correct to say that if a population is finite the sample is a random sample if on each draw each member of the sample has an equal chance of being drawn (this would imply a joint density function of the required form), but this would require sampling with replacement.

Further thought: Consider the condition that **after a draw, each occupant of the urn has an equal chance of being drawn on the next draw**. This condition is necessary for a sample of n draws to be random, but not sufficient.

One could forsake the MGB definition of a random sample, and define random sampling in such a way that is possible to take a random sample from

a finite population when sampling without replacement. For example, one could define a sample as a random sample of size n if it was drawn in such a way that every sample of size n has an equal chance of being drawn. This definition and the MGB definition are equivalent if the population size is infinite or one samples without replacement. With this alternative definition of a random sample, the probability of drawing any sample of size n , with replacement is $m^{-n} = \frac{1}{m^n}$ and, without replacement, $1/(m)_n$. Note that $1/(m)_n \rightarrow m^{-n}$ for finite n as m approaches infinity. This alternative definition of randomness is how Feller defines it. This alternative definition is fine but in the population is finite, one cannot write then joint density function of the "random sample" as the product of identical density functions. Think about this bit more, page 30 Feller

9. Assume that there are n students in our class and that one, and only one, is named Karan. Imagine that we draw, without replacement, a random sample of size of size r from the class.³ What is the probability that Karan appears in the sample? Does not appear in the sample? Now answer these same two questions assuming that one randomly samples with replacement.

answer **without replacement**: Consider first the probability of Karan not appearing in the sample. On the first draw the probability of Karan not appearing is $\frac{n-1}{n}$. For the second draw, the probability of not drawing Karan, assuming Karan is still in the urn, is $\frac{n-2}{n-1}$. For the third draw it is $\frac{n-3}{n-2}$. The product of these for r draws is $\frac{n-r}{n}$,⁴ the probability of not observing Karan in a sample of size r , taken without replacement. And the probability of Karan being in the sample is one minus this, which is $\frac{r}{n}$.

For example, if $n = 32$ and the sample size is 16, Karan has a 50% chance of appearing.

Note there are other ways of writing $\frac{n-r}{n}$ and $\frac{r}{n}$.

For example, in increasing messiness, $\frac{r}{n} = 1 - \frac{n-r}{n} = \frac{(n-1)!}{(r-1)!n!} r!$ Checking this last mess, $\frac{(32-1)!}{(16-1)!32!} 16! = 0.5$, and $\frac{(32-1)!}{(8-1)!32!} 8! = 0.25 = \frac{r}{n} = \frac{8}{32} = 0.25$, so it is probably correct.

For example, $\frac{n-r}{n} = \frac{(n-1)_r}{(n)_r} = \frac{\frac{(n-1)!}{(n-1-r)!}}{\frac{n!}{(n-r)!}} = \frac{(n-1)!}{(n-1-r)!} \left(\frac{(n-r)!}{n!}\right)$ where $(n)_r$ is the number of samples sample of size r taken without replacement from a population of size n , and $(n-1)_r$ is the number of samples sample of size r taken without replacement from a population of size $n-1$ (excluding Karan)

As an aside, and of interest, the probability of Karan on the first draw is $\frac{1}{32} = \frac{1}{n}$, and the probability of him appearing in the sample is $\frac{r}{n}$, as shown above. But $\frac{r}{n} = r\left(\frac{1}{n}\right)$ is the sum of these probabilities (not the product).

³Note that MGB would say one cannot take a random sample from a finite sample if one samples without replacement. So, I being a bit fudgy, here.

⁴For example if $r = 4$, we have $\frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdot \frac{n-3}{n-2} \cdot \frac{n-4}{n-3} = \frac{n-4}{n}$

Note that one can derive the answer by taking the formula in question 1, $\Pr[k] = \frac{\binom{K}{k} \binom{M-K}{n-k}}{\binom{M}{n}}$, noting that here $K = 1$ (one zombie in class, named Karan), $n = r$, the sample size, and $M = n$ equals the population size, and $k = 1$, so $\Pr[1] = \frac{\binom{1}{r-1} \binom{n-1}{n-r}}{\binom{n}{r}} = \frac{\binom{1}{r-1} \binom{n-1}{n-r}}{\binom{n}{r}} = \frac{(n-1)!}{(r-1)!(n-r)!} = \frac{(n-1)!}{(r-1)!n!} = \frac{r!}{(r-1)!n} = \frac{r}{n}$ WOW.

Now consider random sampling **with replacement**. Here I am going to start by determining the probability that Karan does not appear in the sample. Recollect that if one randomly samples with replacement there are n^r samples and each sample has a $\frac{1}{n^r} = \left(\frac{1}{n}\right)^r$ chance of appearing. Further note that $\left(\frac{1}{n}\right)$ is the probability of Karan being chosen on each draw, so $\left(1 - \frac{1}{n}\right)$ is the probability, of him not being drawn on a draw. So, **the probability of him never being drawn is $\left(1 - \frac{1}{n}\right)^r = \left(\frac{n-1}{n}\right)^r$** . In which case, the **probability of him being in the sample one or more times is $1 - \left(1 - \frac{1}{n}\right)^r = 1 - \left(\frac{n-1}{n}\right)^r$** . Note that $\left(1 - \frac{1}{n}\right)^r = \frac{K}{n^r}$ where K is the number of samples that contain one or more Karans. Solving $K = n^r \left(1 - \frac{1}{n}\right)^r$

For example, and without replacement, if $n = 32$ and $r = 16$, there are $n^r = 32^{16} = 1.2089 \times 10^{24}$ samples, the probability of Karan being in the sample, one or more times, is $1 - \left(1 - \frac{1}{n}\right)^r = 1 - \left(1 - \frac{1}{32}\right)^{16} = 0.39829$ and $n^r \left(1 - \frac{1}{n}\right)^r = 32^{16} \left(1 - \frac{1}{32}\right)^{16} = 7.2742 \times 10^{23}$ of the samples contain at least one Karan.