STAT 341 Lab #10
Spring 2022 Tue, Apr 26th Name: Solutions

The Multinomial Distribution

In this lab, we are going to explore a special multivariate discrete distribution called the multinomial distribution. The properties of this distribution are:

- The experiment consists of n independent and identical trials with $k > 2$ possible outcomes on each trial.
- Each possible outcome has probability p_j for $j = 1, 2, \ldots, k$ where $\sum_{i=1}^{n} p_j = 1$.
- The random variables $Y_1, Y_2, \ldots, Y_k$ are the number of trials belonging to the jth outcome in the n trials where $\sum_{j=1}^{k} Y_j = n$.
- The joint pmf is $p(y_1, y_2, \cdots, y_k) = \dfrac{n!}{\prod_{j=1}^{k} y_j!} \prod_{j=1}^{k} p_j^{y_j}$ for $\sum_{j=1}^{k} p_j = 1$ and $\sum_{j=1}^{k} y_j = n$
- R contains a function to calculate multinomial probabilities with the formula above. The function is:

  `dmultinom(y, p)` where `y = (y1, y2, ..., yk)` and `p = (p1, p2, ..., pk)`

The Dirichlet Distribution

In reality, it isn't possible to conduct each trial for the multinomial exactly the same, so if we want a more realistic simulation of reality, we might put a distribution on the probabilities, p_j, to allow them to vary from trial to trial. An appropriate distribution for the p_j's is the Dirichlet distribution. The properties of this distribution are:

- This continuous distribution is the multivariate extension of the beta distribution.
- For $k > 2$ categories, the concentration parameters are $\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_k > 0$
- The random variables $0 < P_1, P_2, \ldots, P_k < 1$ are the probabilities of being in the j^{th} category where for any realization then $\sum_{i=1}^{n} p_j = 1$
- The joint pdf is $f(p_1, p_2, \ldots, p_k) = \dfrac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} p_j^{\alpha_j - 1}$
- R can simulate n Dirichlet random variables using the following function from the `extraDistr` package:

  `rdirichlet(n, alpha)` where `alpha = (a1, a2, ..., ak)`

Activity 1: Simplified Dice - Suppose you have a 10-sided dice with sides numbered 0 through 9 with each side being *exactly* equally likely. Imagine rolling this dice 100 times and counting the number of rolls for each outcome 0 through 9.

1. Write the joint probability mass function for the random variables Y_j = the number of rolls for the jth outcome, for $j = 1, 2, \ldots, 10$.

   The joint pmf is $p(y_1, y_2, \cdots, y_{10}) = \dfrac{100!}{\prod_{j=1}^{10} y_j!} \prod_{j=1}^{10} \dfrac{1}{10}^{y_j}$ for $\sum_{j=1}^{10} y_j = 100$

2. Find the probability each random variable Y_j is equal to 10, for $j = 1, 2, \ldots, 10$.

   `dmultinom(x=rep(10,10), prob=rep(1/6, 10))` $= 2.357075 \times 10^{-8}$

3. Find the probability of obtaining the following outcome for the random variables $Y = (8, 12, 9, 10, 10, 8, 15, 13, 7, 8)$.

   `dmultinom(x=c(8, 12, 9, 10, 10, 8, 15, 13, 7, 8), prob=rep(1/6, 10))` $= 1.515673 \times 10^{-9}$

4. Simulate repeating the experiment (i.e. rolling the dice 100 times and counting the number of times each side is rolled) $M = 20$ times using the function `rmultinom(n, size, prob)` where `n` $= M$, `size` $= \sum y_j$, and `prob=rep(1/6,10)`. Then, use R to estimate the means, variances, and covariances.

```
p1 = rep(1/6, 10)
Ys = rmultinom(n=20, size=100, prob=p1)
rowMeans(Ys)
apply(Ys, 1, var)
cov(t(Ys))
```

```
> #### Activity 1 ####
> p1 = rep(1/6, 10)
> # Q4
> Ys = rmultinom(n=20, size=100, prob=p1)
> rowMeans(Ys)
 [1] 10.55 10.90  8.80 10.65 10.20  9.90 10.10 10.05  8.95  9.90
> apply(Ys, 1, var)
 [1]  6.786842 11.357895  7.852632  9.397368 12.273684  8.200000 10.200000  7.944737  4.576316  8.831579
> cov(t(Ys))
           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]       [,8]       [,9]      [,10]
 [1,]  6.7868421 -4.73157895 -1.30526316  1.83421053 -0.1684211 -2.5736842  0.2052632 -2.1342105 -0.65526316  2.7421053
 [2,] -4.7315789 11.35789474 -0.86315789 -1.87894737  2.2315789  1.1473684 -3.7263158 -1.9947368 -0.05789474 -1.4842105
 [3,] -1.3052632 -0.86315789  7.85263158 -0.02105263 -3.8000000 -0.7052632  3.0210526  1.1157895 -1.95789474 -3.3368421
 [4,]  1.8342105 -1.87894737 -0.02105263  9.39736842 -4.0842105 -2.6684211  2.2473684 -3.0342105 -3.96578947  2.1736842
 [5,] -0.1684211  2.23157895 -3.80000000 -4.08421053 12.2736842  1.7578947 -4.5473684 -1.6947368  1.27368421 -3.2421053
 [6,] -2.5736842  1.14736842 -0.70526316 -2.66842105  1.7578947  8.2000000 -2.8315789 -1.1526316  1.36315789 -2.5368421
 [7,]  0.2052632 -3.72631579  3.02105263  2.24736842 -4.5473684 -2.8315789 10.2000000 -0.7421053 -0.99473684 -2.8315789
 [8,] -2.1342105 -1.99473684  1.11578947 -3.03421053 -1.6947368 -1.1526316 -0.7421053  7.9447368  1.21315789  0.4789474
 [9,] -0.6552632 -0.05789474 -1.95789474 -3.96578947  1.2736842  1.3631579 -0.9947368  1.2131579  4.57631579 -0.7947368
[10,]  2.7421053 -1.48421053 -3.33684211  2.17368421 -3.2421053 -2.5368421 -2.8315789  0.4789474 -0.79473684  8.8315789
```

Activity 2: Mandel's Peas - In early research on genetics and inherited traits, Mandel crossed peas with color traits (yellow and green) and shape traits (round and wrinkled) to observe dominate (yellow and round) and recessive (green and wrinkled) traits. After two generations, Mandel observed the probabilities of obtaining:

- yellow and round peas = 9/16
- yellow and wrinkled peas = 3/16
- green and round peas = 3/16
- green and wrinkled peas = 1/16

Suppose you have 400 second generation plants and determine the color and shape traits of the peas from each plant.

5. Write the joint probability density function for the random variables Y_j = the number of plants for each combination of color and shape traits, for $j = 1, 2, 3, 4$.

   The joint pmf is

   $$p(y_1, y_2, \cdots, y_k) = \frac{400!}{y_1!y_2!y_3!y_4!} \left(\frac{9}{16}\right)^{y_1} \left(\frac{3}{16}\right)^{y_2} \left(\frac{3}{16}\right)^{y_3} \left(\frac{1}{16}\right)^{y_4}$$

   for $\sum_{j=1}^{4} y_j = 400$

6. Find the probability of obtaining the following outcome for the random variables $Y = (225, 75, 75, 25)$.

   `dmultinom(x=c(225, 75, 75, 25), prob=c(9, 3, 3, 1)/16)` $= 0.0002244681$

7. Find the probability of obtaining the following outcome for the random variables $Y = (247, 53, 85, 15)$.

   `dmultinom(x=c(247, 53, 85, 15), prob=c(9, 3, 3, 1)/16)`
   $= 1.516464 \times 10^{-7}$

Activity 3: Realistic Dice - Suppose you have a 10-sided dice with sides numbered 0 through 9 and in reality you can't roll the dice exactly the same way each time. Simulate rolling this dice 100 times, each roll with a unique vector of multinomial probabilities simulated from the Dirichlet distribution and then counting the number of rolls for each outcome 0 through 9.

8. If you feel like the probability of each side is *almost* equally likely, then you might use a concentration vector for the Dirichlet distribution of $\alpha_j = 1$ for all $j = 1, 2, \ldots, 10$. Simulate 100 different rolls and summarize the number of rolls corresponding to each of the 10 outcomes (i.e. find the multinomial outcome).

```
a1 = rep(1,10)
ps_equal = rdirichlet(n=100, alpha=a1)
x1 = rowSums(apply(ps_equal, 1, function(x)
{       rmultinom(n=1, size=1, prob=x)       }))
```

$$\text{x1} = (10, 7, 13, 6, 13, 11, 7, 9, 9, 15)$$

9. Find the probability of obtaining the result from Exercise 7 assuming the probabilities of obtaining each side are equally likely.

`dmultinom(x=x1, prob=rep(1/6, 10))` $= 5.276003 \times 10^{-10}$

10. If you feel like the probability of rolling bigger numbers is more likely than rolling smaller numbers (for example, $p_1 < p_2 < \ldots < p_9 < p_{10}$, then you might use a concentration vector for the Dirichlet distribution of $\alpha_j = j$ for all $j = 1, 2, \ldots, 10$. Simulate 100 different rolls and summarize the number of rolls corresponding to each of the 10 outcomes (i.e. find the multinomial outcome).

```
a2 = seq(1,10,1)
ps_biased = rdirichlet(n=100, alpha=a2)
x2 = rowSums(apply(ps_biased, 1, function(x)
{       rmultinom(n=1, size=1, prob=x)       }))
```

$$\text{x2} = (4, 4, 9, 3, 9, 9, 16, 22, 10, 14)$$

11. Find the probability of obtaining the result from Exercise 9 assuming the probabilities of obtaining each side are equally likely. How does this compare to your answer from Exercise 8?

`dmultinom(x=x2, prob=rep(1/6, 10))` $= 7.59597e \times 10^{-15}$