## Announcements

- Make sure to sign in on the google form (I send a list of what section questions are useful for what pset questions afterwards)
- Pset 2 due Friday 2/10

## Alphabet Soup

The following question deals with the set of four letter words in the Scrabble dictionary available here.

In this question, we will be modeling the letters in four letter words as draws from a multinomial distribution: $\vec{Y} \sim \text{Mult}_k(4, \vec{p})$. Recall that the Multinomial PMF is

$$P(Y_1 = n_1, ..., Y_k = n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i}$$

1. What should $k$ be?
2. Suppose we generated a draw from the distribution. What additional step would we have to perform to construct a word from the draw?
3. What assumption of the Multinomial distribution is obviously violated here?
4. Find the likelihood function $L(\vec{p}; \mathbf{Y})$ where $\mathbf{Y}$ is a $n \times k$ matrix with one row for each word and one column for each letter. (E.g. "face" would become a row $1, 0, 1, 0, 1, 1, 0, ..., 0.$) Then, find the log likelihood function. What constants can we drop?

5. If we maximized the log likelihood as it stands now, we would end up with $p_j = \infty$ for all $p_j$. To prevent this, we'll restrict the sum of the $p_j$ with a Lagrangian constraint. Specifically, find the gradient of $\ell(\vec{p}; \vec{y}) + \lambda(1 - \sum_{j=1}^{k} p_j)$ (the derivative with respect to each $p_j$), set it equal to 0, solve for $p_j$, and then use the fact that $\sum_{j=1}^{k} p_j = 1$ to solve for $\lambda$. Explain in words what this MLE is.

6. Find the $\hat{p_j}$ from the data and display them visually. (The `strsplit`, `unlist`, and `table` functions may be useful.)

```
# TODO: Make a data frame with one column of letters and the other column of proportions

df <- data.frame(matrix(nrow = 0, ncol = 2))
colnames(df) <- c("letters", "proportions")

ggplot(df, aes(x = toupper(letters), y = proportions)) +
  geom_bar(stat="identity") +
  xlab("Letter") +
  ylab("Proportion") +
  theme_bw()
```

7. Suppose you generate a set of 4 letters from the multinomial distribution above and put the 4 letters in a random order. What is the probability of producing the word "stat"? Find this in two ways: first, condition on the multinomial. Second, use counting. Then, find this probability in the data set.

```
# TODO: Calculate the probability for this data set
```

8. With the magic of seed setting, we can make this probability 1! (Interestingly, the probability above means the expected number of seeds until generating "stat" is 43478, and the actual seed was 31083.)

```
set.seed(31083)

# Make a single draw from Mult(4, p)
draw <- as.vector(rmultinom(1, 4, df$proportions))
```

```r
# Turn this draw into a vector of letters
bag_of_letters <- vector()
for (i in 1:length(letters)) {
  bag_of_letters <- c(bag_of_letters, rep(letters[i], draw[i]))
}

# Put these letters in a random order
word <- paste0(sample(bag_of_letters, 4), collapse = "")

print(word)
```

9. Find a maximum likelihood estimator for $\text{Cov}(\vec{Y}_{[1]}, \vec{Y}_{[2]})$, the covariance between the number of As and the number of Bs.

10. Find a method of moments estimator for the covariance between the number of As and the number of Bs.

11. Compare the MSEs of these two estimators through simulation with $n = 30$ "words" in each draw.

```r
nsims <- 10000
n <- 30

single_sim <- function() {
  # TODO: Make n draws from a multinomial, calculate the MLE, calculate the MOM

  mle <- NA

  mom <- NA

  return (c(mle, mom))
}

# Run this nsims times, storing results in a matrix
estimates <- replicate(nsims, single_sim())

# TODO: Get the true covariance
```

```
true_cov <- NA

# MLE MSE
print(mean((estimates[1,] - true_cov)^2))
```

```
## [1] NA
```

```
# MOM MSE
print(mean((estimates[2,] - true_cov)^2))
```

```
## [1] NA
```

## Logistic Logic

The Logistic($s$) distribution is defined to be the distribution of $s \log\left(\frac{U}{1-U}\right)$ where $U \sim \text{Unif}(0,1)$.

1. Find the CDF $F(y)$ of the Logistic distribution. (Hint: Let $Y$ have the Logistic distribution and let $U$ have the Uniform distribution. Then, write $Y$ in terms of $U$, isolate $U$, and use the Uniform PDF.)

2. Find the PDF of the Logistic distribution.

3. Let $Y$ have the Logistic distribution. Find $E(Y)$ and $\text{Var}(Y)$. You may use the facts that $\int_0^1 \ln\left(\frac{x}{1-x}\right) dx = 0$ and $\int_0^1 \ln^2\left(\frac{x}{1-x}\right) dx = \pi^2/3$.

4. Suppose we are measuring vehicle velocities on a congested highway which are distributed Logistic($s$). (It makes sense for the distribution to be symmetric around 0 since the vehicles are equally likely to be going either direction). However, our instrument can only measure velocities in the range of $[-c, c]$ (for any physicists in the room, $c$ does not represent the speed of light). We want to estimate $s$ despite this limitation. Find the likelihood function for $s$ given $n_1$ observed velocities $Y_1, ..., Y_{n_1}$, $n_2$ velocities less than $-c$ and $n_3$ velocities more than $c$.

5. A closed form solution for the maximum likelihood estimator of $s$ does not exist (or at least I wasn't able to find it after an hour of number pushing, and Google doesn't seem to have it either). Instead, consider the following method of moments estimator for $s$:

$$\hat{s} = \sqrt{\frac{3}{n_1 \pi^2} \sum_{i=1}^{n_1} Y_i^2}$$

Describe the logic behind this estimator.

6. Find the sign of the bias of $\hat{s}$ for $s$ with an argument about how $E\left(\frac{3}{n\pi^2} \sum_{i=1}^{n_1} Y_i^2\right)$ compares to $s^2$ and Jensen's inequality. (Since the square root function is concave, Jensen's inequality says $E(\sqrt{X}) < \sqrt{E(X)}$.)