What if the respondents in the survey had three choices: (1) I feel optimistic, (2) I don’t feel optimistic, (3) I am not sure? What would you consider as an appropriate sampling distribution?
The Multinomial Distribution

Origins
The multinomial distribution arises from an extension of the binomial experiment to situations where each trial has $k \geq 2$ possible outcomes.

Suppose that we have an experiment with

- $n$ independent trials, where
- each trial produces exactly one of the events $E_1, E_2, \ldots, E_k$

  (i.e. these events are mutually exclusive and collectively exhaustive), and

- on each trial, $E_j$ occurs with probability $\pi_j$, $j = 1, 2, \ldots, k$.

Notice that

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1.$$
Define the random variables

\[ X_1 = \text{number of trials in which } E_1 \text{ occurs}, \]
\[ X_2 = \text{number of trials in which } E_2 \text{ occurs}, \]
\[ \vdots \]
\[ X_k = \text{number of trials in which } E_k \text{ occurs}. \]

Then \( X = (X_1, X_2, \ldots, X_k) \) is said to have a multinomial distribution with index \( n \) and parameter \( \pi = (\pi_1, \pi_2, \ldots, \pi_k) \). In most problems, \( n \) is regarded as fixed and known.

The individual components of a multinomial random vector are binomial,

\[ X_1 \sim \text{Bin}(n, \pi_1), \]
\[ X_2 \sim \text{Bin}(n, \pi_2), \]
\[ \vdots \]
\[ X_k \sim \text{Bin}(n, \pi_k). \]

However, the components are not independent. Even though the individual \( X_j \)'s are random, their sum

\[ X_1 + X_2 + \cdots + X_k = n \]

is fixed. Therefore, the \( X_j \)'s are negatively correlated.
Notation. If $X = (X_1, X_2, \ldots, X_k)$ is multinomially distributed with index $n$ and parameter $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$, then we will write

$$X \sim \text{Mult}(n, \pi).$$

Distribution function. The probability that $X = (X_1, \ldots, X_k)$ takes a particular value $x = (x_1, \ldots, x_k)$ is

$$f(x) = \frac{n!}{x_1! x_2! \cdots x_k!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k}.$$

The possible values of $X$ are the set of $x$-vectors such that each $x_j \in \{0, 1, \ldots, n\}$ and $x_1 + \cdots + x_k = n.$
Example. Suppose that the racial/ethnic distribution in a large city is given by this table:

<table>
<thead>
<tr>
<th></th>
<th>Black</th>
<th>Hispanic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20%</td>
<td>15%</td>
<td>65%</td>
</tr>
</tbody>
</table>

Suppose that a jury of twelve members is chosen from this city in such a way that each resident has an equal probability of being selected independently of every other resident.
Let’s find probability that the jury contains

- three Black, two Hispanic, and seven Other members;
- four Black and eight Other members;
- at most one Black member.

To solve this problem, let $X = (X_1, X_2, X_3)$ where $X_1$ = number of Black members, $X_2$ = number of Hispanic members, and $X_3$ = number of Other members. Then $X$ has a multinomial distribution with parameters $n = 12$ and $\pi = (.20, .15, .65)$. The answer to the first part is

$$P(X_1 = 3, X_2 = 2, X_3 = 7) = \frac{n!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3} = \frac{12!}{3! 2! 7!} (.20)^3 (.15)^2 (.65)^7 = 0.0699.$$  

The answer to the second part is

$$P(X_1 = 4, X_2 = 0, X_3 = 8) = \frac{12!}{4! 0! 8!} (.20)^4 (.15)^0 (.65)^8 = 0.0252.$$
For the last part, note that “at most one Black member” means $X_1 = 0$ or $X_1 = 1$. $X_1$ is a binomial random variable with $n = 12$ and $p = \pi_1 = .2$. Using the binomial probability distribution,

$$
P(X_1 = 0) = \frac{12!}{0!12!} (.2)^0 (.8)^{12}
\quad = 0.0687
$$

and

$$
P(X_1 = 1) = \frac{12!}{1!11!} (.2)^1 (.8)^{11}
\quad = 0.2061.
$$

Therefore, the answer is

$$
P(X_1 = 0) + P(X_1 = 1) = 0.0687 + 0.2061 = 0.2748.
$$
Moments. Many of the elementary properties of the multinomial can be derived by decomposing $X$ as the sum of iid random vectors

$$X = Y_1 + Y_2 + \cdots Y_n,$$

where each $Y_i \sim \text{Mult}(1, \pi)$. In this decomposition, $Y_i$ represents the outcome of the $i$th trial; it’s a vector with a 1 in position $j$ if $E_j$ occurred on that trial and 0’s in all other positions. The elements of $Y_i$ are correlated Bernoulli’s. For example, with $k = 2$ possible outcomes on each trial, the possible values of $Y_i$ are

$$\begin{align*}
(1, 0) & \quad \text{with probability } \pi_1, \\
(0, 1) & \quad \text{with probability } \pi_2 = 1 - \pi_1.
\end{align*}$$

Because the individual elements of $Y_i$ are Bernoulli, the mean of $Y_i$ is $\pi = (\pi_1, \pi_2)$, and its covariance matrix is

$$
\begin{bmatrix}
\pi_1 (1 - \pi_1) & -\pi_1 \pi_2 \\
-\pi_1 \pi_2 & \pi_2 (1 - \pi_2)
\end{bmatrix}.
$$

(To see that the off-diagonal elements are $-\pi_1 \pi_2$, one can use the definition of covariance in Lecture 1.)
More generally, with $k$ possible outcomes, the mean of $Y_i$ is $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$, and the covariance matrix is

$$
\begin{bmatrix}
\pi_1(1 - \pi_1) & -\pi_1 \pi_2 & \cdots & -\pi_1 \pi_k \\
-\pi_1 \pi_2 & \pi_2(1 - \pi_2) & \cdots & -\pi_2 \pi_k \\
\vdots & \vdots & \ddots & \vdots \\
-\pi_1 \pi_k & -\pi_2 \pi_k & \cdots & \pi_k(1 - \pi_k)
\end{bmatrix}.
$$
Using the fact that 

\[ X = Y_1 + Y_2 + \cdots + Y_n, \]

it immediately follows that the mean of \( X \) is 

\[ n\pi = (n\pi_1, n\pi_2, \ldots, n\pi_k) \]

and the covariance matrix for \( X \) is

\[
\begin{bmatrix}
  n\pi_1(1 - \pi_1) & -n\pi_1\pi_2 & \cdots & -n\pi_1\pi_k \\
n\pi_1\pi_2 & n\pi_2(1 - \pi_2) & \cdots & -n\pi_2\pi_k \\
\vdots & \vdots & \ddots & \vdots \\
-n\pi_1\pi_k & -n\pi_2\pi_k & \cdots & n\pi_k(1 - \pi_k)
\end{bmatrix}
\]

Because the elements of \( X \) are constrained to sum to \( n \), this covariance matrix is singular. If all the \( \pi_j \)'s are positive, then the covariance matrix has rank \( k - 1 \). Intuitively, this makes sense; the last element \( X_k \) can be replaced by \( n - X_1 - X_2 - \cdots - X_{k-1} \); there are really only \( k - 1 \) “free” elements in \( X \). If some elements of \( \pi \) are zero, the rank drops by one for every zero element.
Parameter space. If we don’t impose any restrictions on the parameter

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k)$$

other than the logically necessary constraints

$$\pi_j \in [0, 1], \ j = 1, \ldots, k \quad (1)$$

and

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1, \quad (2)$$

then the parameter space is the set of all $\pi$-vectors that satisfy (1) and (2). This set is called a simplex. In the special case of $k = 3$, we can visualize $\pi = (\pi_1, \pi_2, \pi_3)$ as a point in three-dimensional space. The simplex $\mathcal{S}$ is the triangular portion of a plane with vertices at $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$:
More generally, the simplex is a portion of a $(k - 1)$-dimensional hyperplane in $k$-dimensional space. Alternatively, we can replace

$$\pi_k$$

by $1 - \pi_1 - \pi_2 - \cdots - \pi_{k-1}$

because it’s not really a free parameter and view the simplex in $(k - 1)$-dimensional space. For example, with $k = 3$, we can replace $\pi_3$ by $1 - \pi_1 - \pi_2$ and view the parameter space as a triangle:

ML estimation. If $X = \text{Mult}(n, \pi)$ and we observe $X = x$, then the loglikelihood function for $\pi$ is

$$l(\pi; x) = x_1 \log \pi_1 + x_2 \log \pi_2 + \cdots + x_k \log \pi_k.$$
Using multivariate calculus, it’s easy to maximize this function subject to the constraint
\[ \pi_1 + \pi_2 + \cdots + \pi_k = 1; \]
the maximum is achieved at
\[ p = n^{-1} x \]
\[ = (x_1/n, x_2/n, \ldots, x_k/n), \]
the vector of sample proportions. The ML estimate for any individual \( \pi_j \) is \( p_j = x_j/n \), and an approximate 95% confidence interval for \( \pi_j \) is
\[ p_j \pm 1.96 \sqrt{\frac{p_j(1-p_j)}{n}} \]
because \( X_j \sim \text{Bin}(n, \pi_j) \).
Fusing cells. We can collapse a multinomial vector by fusing cells (i.e. by adding some of the cell counts $X_j$ together). If

$$X = (X_1, X_2, \ldots, X_k) \sim \text{Mult}(n, \pi)$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$, then

$$X^* = (X_1 + X_2, X_3, X_4, \ldots, X_k)$$

is also multinomial with the same index $n$ and modified parameter

$$\pi^* = (\pi_1 + \pi_2, \pi_3, \pi_4, \ldots, \pi_k).$$

In the multinomial experiment, we are simply fusing the events $E_1$ and $E_2$ into the single event “$E_1$ or $E_2$.” Because these events are mutually exclusive,

$$P(E_1 \text{ or } E_2) = P(E_1) +, P(E_2) = \pi_1 + \pi_2.$$
Partitioning the multinomial. We can also partition the multinomial by conditioning on (treating as fixed) the totals of subsets of cells. For example, consider the conditional distribution of $X$ given that

\[ X_1 + X_2 = z, \]
\[ X_3 + X_4 + \cdots + X_k = n - z. \]

The subvectors $(X_1, X_2)$ and $(X_3, X_4, \ldots, X_k)$ are conditionally independent and multinomial,

\[
(X_1, X_2) \sim \text{Mult} \left[ z, \left( \frac{\pi_1}{\pi_1 + \pi_2}, \frac{\pi_2}{\pi_1 + \pi_2} \right) \right], \\
(X_3, \ldots, X_k) \sim \text{Mult} \left[ n - z, \left( \frac{\pi_3}{\pi_3 + \cdots + \pi_k}, \ldots, \frac{\pi_k}{\pi_3 + \cdots + \pi_k} \right) \right].
\]

The joint distribution of two or more independent multinomials is called “product-multinomial.” If we condition on the sums of non-overlapping groups of cells of a multinomial vector, it’s distribution splits into product multinomial. The parameter for each part of the product multinomial is a portion of the original $\pi$ vector, normalized to sum to one.
Relationship between the multinomial and Poisson.
Suppose that $X_1, X_2, \ldots, X_k$ are independent Poisson random variables,

\[ X_1 \sim P(\lambda_1), \]
\[ X_2 \sim P(\lambda_2), \]
\[ \vdots \]
\[ X_k \sim P(\lambda_k), \]

where the $\lambda_j$'s are not necessarily equal. Then the conditional distribution of the vector

\[ X = (X_1, X_2, \ldots, X_k) \]

given the total

\[ n = X_1 + X_2 + \cdots + X_k \]
is Mult$(n, \pi)$, where

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k)$$

and

$$\pi_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_k}.$$ 

That is, $\pi$ is simply the vector of $\lambda_j$’s normalized to sum to one.

This fact is important, because it implies that the unconditional distribution of $(X_1, \ldots, X_k)$ can be factored into the product of two distributions: a Poisson distribution for the overall total,

$$n \sim P(\lambda_1 + \lambda_2 + \cdots + \lambda_k),$$

and a multinomial distribution for

$X = (X_1, X_2, \ldots, X_k)$ given $n$,

$$X \sim \text{Mult}(n, \pi).$$

The likelihood factors into two independent functions, one for $\sum_{j=1}^{k} \lambda_j$ and the other for $\pi$. The total $n$ carries no information about $\pi$ and vice-versa.
Therefore, likelihood-based inferences about \( \pi \) are the same whether we regard \( X_1, \ldots, X_k \) as sampled from \( k \) independent Poissons or from a single multinomial. That is, any estimates, tests, etc. for \( \pi \) or functions of \( \pi \) will be the same whether we regard \( n \) as random or fixed.

**Example.** Suppose that you wait at a busy intersection for one hour and record the color of each vehicle as it drives by. Let

\[
\begin{align*}
X_1 & = \text{number of white vehicles} \\
X_2 & = \text{number of black vehicles} \\
X_3 & = \text{number of silver vehicles} \\
X_4 & = \text{number of red vehicles} \\
X_5 & = \text{number of blue vehicles} \\
X_6 & = \text{number of green vehicles} \\
X_7 & = \text{number of any other color}
\end{align*}
\]

In this experiment, the total number of vehicles observed,

\[
n = X_1 + X_2 + \cdots + X_7,
\]
is random. (It would have been fixed if, for example, we had decided to classify the first \( n = 500 \) vehicles we see. But because we decided to wait for one hour, the \( n \) is random.)

In this case, it’s reasonable to regard the \( X_j \)’s as independent Poisson random variables with means \( \lambda_1, \lambda_2, \ldots, \lambda_7 \). But if our interest lies not in the \( \lambda_j \)’s but in the proportions of various colors in the vehicle population, inferences about these proportions will be the same whether we regard the sample size \( n \) as random or fixed. That is, we can proceed as if

\[
X = (X_1, \ldots, X_7) \sim \text{Mult}(n, \pi)
\]

where \( \pi = (\pi_1, \ldots, \pi_7) \), even though \( n \) is actually random.

**Next time:** Goodness of fit testing for one-way tables