Analysis of Categorical Data

(Three or More Outcomes)

A multinomial experiment

1. The experiment consists of n identical trials.
2. The outcome of each trial falls into one of k cells.
3. The probability of an outcome falling into cell i is π_i and remains the same from trial to trial.
4. The trials are independent.

Suppose we toss a die 600 times. Is this an example of a multinomial experiment? If not, why not? If it is, then what are n, k, and the π_i's?

Genes can be either dominant (represented by capital letters) or recessive (represented by a small letter). Two characteristics of pea plants are their color and texture. Yellow (Y) is the dominant color, while green (y) is recessive. For texture, round (R) is dominant and wrinkled (r) is recessive. Below is a table which shows the possible offspring of two hybrid (Yy,Rr) pea plants.

	(YR)	(Yr)	(yR)	(yr)
(YR)				
(Yr)				
(yR)				
(yr)				

What proportion should be yellow and round? Yellow and wrinkled? Green and round? Green and wrinkled?

Is this an example of a multinomial experiment? If so, what are k and the π_i's?

Hypothesis Testing for a One-way Frequency Table

Hypotheses

H_0: $\pi_1=\pi_{10}, \pi_2=\pi_{20}, \ldots, \pi_k=\pi_{k0}$
H_a: not all the above true

Test Statistic:

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}\ ,\ \ where\!: n_i = the\ number\ of\ \ trials\ with\ outcome\ i$$

$$E_i = n\pi_i$$

Distribution for the Rejection Region:

The rejection region is the upper tail of a chi-square distribution for α with k-1 degrees of freedom.

Assumptions:

1. The observed values are the result of a multinomial experiment.
2. The expected value for each cell should be 5 or greater.

Suppose that the results of 600 die tosses revealed that the number 1 was observed 110 times, 2 was observed 90 times, 3 was observed 120 times, 4 was observed 80 times, 5 was observed 100 times, and 6 was observed 100 times. Is there sufficient evidence to conclude that the die is "loaded"?

Suppose that 160 peas from the offspring of a pea plant which was cross-fertilized revealed that 85 were yellow and round, 30 were yellow and wrinkled, 28 were green and round, and 17 were green and wrinkled. Do these results provide sufficient evidence to conclude that both parents weren't hybrids?

Two-Way Frequency Tables

There are two types of two-way frequency tables that can occur.

1. Contingency tables
2. Fixed row or column tables

Contingency Tables

The 2-dimensional contingency table examines the relationship between two variables to determine if one variable is **contingent** (dependent) on the value of the second variable.

In such situations, one chooses a fixed total sample size and then places each observation in the appropriate category. (This is one multinomial experiment.)

Example: A pollster telephones 1000 people and asks each participant two questions: (1) Are you a republican, a democrat, or an independent? (2) Who did you vote for in the past election? The pollster wants to determine if the voter's response is contingent on their party affiliation.

	Republican	Democrat	Independent
Bush			
Gore			
Other			

Fixed Row or Column Tables

The fixed row or column table examines the discrete distributions for each value of a variable to determine if the distributions are different.

In such situations, one chooses fixed sample sizes for each value of a variable and then places each observation in the appropriate category for the discrete distribution. (This is a series of multinomial experiments.)

Example: A researcher wants to determine if the antibiotic Tetracycline helps acne patients. To do this, 100 patients are given the drug, and another 100 patients are given a placebo. The researcher then looks at the number reporting improvement for each group to determine if there is a difference in the proportion improved.

	Improved	Not-Improved
Placebo		
Tetracycline		

Hypothesis Testing for Two-way Frequency Tables

Hypotheses

For Dependence of Variables X and Y
H_0: Variables X and Y are independent
H_a: Variables X and Y are dependent

For Non-homogeneity
H_0: Distributions for each level of X are homogeneous
H_a: Distributions for each level of X are non-homogeneous

Test Statistic:

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

$$\begin{aligned} where\text{:}\ n_{ij} &= the\ observed\ number\ of\ outcomes\ for\ cell\ ij \\ E_{ij} &= \frac{R_i C_j}{n} \\ R_i &= the\ sum\ of\ Row\ i \\ C_j &= the\ sum\ of\ Column\ j \end{aligned}$$

Distribution for the Rejection Region

The rejection region is the upper tail of a chi-square distribution for α with (r-1)(c-1) degrees of freedom (where r is the number of rows and c is the number of columns).

Assumptions

1. The observed values are the result of one or more multinomial experiments.
2. The expected value for each cell should be 5 or greater.

Example

Students withdraw from college for a variety of reasons. It has been conjectured that the withdrawal rate for students in a university “Freshman Experience” special housing program is different from that for freshman in regular dorms. To test this conjecture a researcher looked at the withdrawal behavior of 500 regular housing freshmen and 400 Freshman Experience students. Below are the results of the study.

	Withdrew from College	
Housing	Yes	No
Freshman Exp.	20	380
Regular	47	453

Does the data provide sufficient evidence to support the researcher's conjecture? (Use a level of significance of 5%.)

Example

Four hundred criminal cases were examined to determine if there is a relationship between the type of offense and the disposition. Below are the results of the cases examined for the study.

	Disposition		
Type of Offense	Prison (Over 5 years)	Prison (5 years or less)	No Prison
Violent	86	24	10
Non-violent	44	150	86

Do the data provide sufficient evidence that disposition is dependent on the type of offense? (Use a significance level of 5%.)

E-11=120*130/400=39
E-12=120*174/400=52.2
E-13=120*96/400=230.4
E-21=280*130/400=91
E-22=280*174/400=121.8
E-23=280*96/400=67.2

Test statistic=(86-39)^2/39+(24-52.2)^2/52.2+(10-230.4)^2/230.4+(44-91)^2/91+(150-121.8)^2/121.8+(86-67.2)^2/67.2=