Packet 7: More Testing

Chap 9.2 Contigency Tables

A contingency table is a type of table that displays the distribution of the combination of multiple categorical variables. For instance, y_{ij} is the frequency of obsering ith category of the first variable and jth category of the second variable.

Multinomial Sampling: $Y \sim Multi(n, p)$, where $n = \sum_i \sum_j Y_{ij}$, $p = \{p_{ij}\}$.

We will learn

1. Homogeneity Test: Test whether two multinomial distributions are equal (each has k categories).

$$H_0 : p_{11} = p_{21}, p_{12} = p_{22} \ldots, p_{1k} = p_{2k}$$

Extend to $H_0 : F(x) = G(y)$.

2. Independence Test: Test for independence of two multinomial random variables.

$$H_0 : P(A_i, B_j) = P(A_i)P(B_j)$$

for all i and j.

Extend to H_0: X and Y independent.

Example 1: A university admissions officers were concerned that males and females students applying for the four different schools (business, engineering, liberal arts, and science) at her university. They collected the following data on the acceptance of 1200 males and 800 females who applied to the university.

Accepted	Business School	Engineering School	Liberal Arts School	School of Science	Total
Male	240	480	120	360	1200
Female	240	80	320	160	800
Total	480	560	440	520	2000

The number in the above table denotes the number falling into each combination of row and column attributes.

Are males and females distributed equally among the various schools?

Accepted	Business School	Engineering School	Liberal Arts School	School of Science	Total
Male	240	480	120	360	1200
Female	240	80	320	160	800
Total	480	560	440	520	2000

Let $p_{\cdot j}$ be the probability of entering school j.

Homogeneity Test: $H_0 : p_{1j} = p_{2j}$ for $j = 1, 2, 3, 4$.

Accepted	Business School	Engineering School	Liberal Arts School	School of Science	Total
Male	240	480	120	360	1200
Female	240	80	320	160	800
Total	480	560	440	520	2000

Independence Test: Whether gender and major are independent?

$H_0 : p_{ij} = p_{i.} \times p_{.j}$ for all i and j.

For continous random variables, we can still apply chi-square test after discretizing the continous random variable into k categries.

Example 2: We measure IQ from 100 random selected individuals and would like to know whether IQ follows $N(100, 16^2)$.

With this idea, we can test

1. Whether X follows a continous distribution.
2. Whether X and Y follows the same distribution.
3. Whether X and Y are independent.

More Examples: 9.2-1, 9.2-2, 9.2-3, 9.2-4

Capture – Recapture Study

The goal is to estimate unknown population size. E.g., how many fish are there in a lake?

Capture: Typically a researcher visits a study area and uses traps to capture a group of individuals alive. Each of these individuals is marked with a unique identifier (e.g., a numbered tag or band), and then is released unharmed back into the environment. Sufficient time is allowed to pass for the marked individuals to redistribute themselves among the unmarked population.

Recapture: Next, the researcher returns and captures another sample of individuals. Then the second sample contains both marked and unmarked individuals.

	Recapture Yes	Recapture Yes
Capture Yes	a	b
Capture No	c	d

$a + b$ = total number of fished marked, $a + c$ = total number of fishes recaptured. $N = a + b + c + d$ is the population size, and d is unknown.

Assumptions:

1. Closed population: no individual dies or is born or move into/out the study area between the two visits. So the waiting time can not be too long, usually 1 week.
2. Capture $\perp$ Recapture (independent).

$$P(captured) = \frac{a+b}{N} = \frac{a}{a+c} = P(captured|recaptured)$$

$$\rightarrow \hat{N} = \frac{(a+b)(a+c)}{a}$$

which is known as Lincoln-Peterson estimater.