

Module 4E: Hypothesis Testing

Revisiting Quantitative evidence & uncertainty

Agenda:

- Poisson Distribution of the four steps in Hypothesis Testing
 1. H_0/H_A : Our model of the test universe (the distribution of the variable)
 2. Test & assumptions: are the assumptions met? Is the test valid?
 3. Quantitative evidence: **p-value**, or critical value.
 4. Conclusion & uncertainty/estimation
- Fisher's Exact Test (McDonald-Kreitman)

Contingency Analysis

Contingency: allows us to determine if two categorical variables are associated (some contingency tests will allow us to quantify the degree of association as well, but not all do this).

Major tests:

- **χ^2 Contingency Test** → similar but not exactly as the same χ^2 Goodness of fit test. You can think of it as a subset of χ^2 Goodness of fit tests with some calculation differences. Basis of test is Multiplication rule with the assumption of independence. Degrees of freedom are calculated differently!
- **Odds ratio** → H_0 : OR=1. Challenge: transforming the sampling distribution of OR so that it is normally distributed.
- **Relative Risk** → like OR but accounts for proportion of (rare) event in the population
- **Fisher's Exact test** → exact calculation. You can think of it as the contingency version of calculating a p-value

Contingency Analysis:

Review prompt: Associations between categorical variables

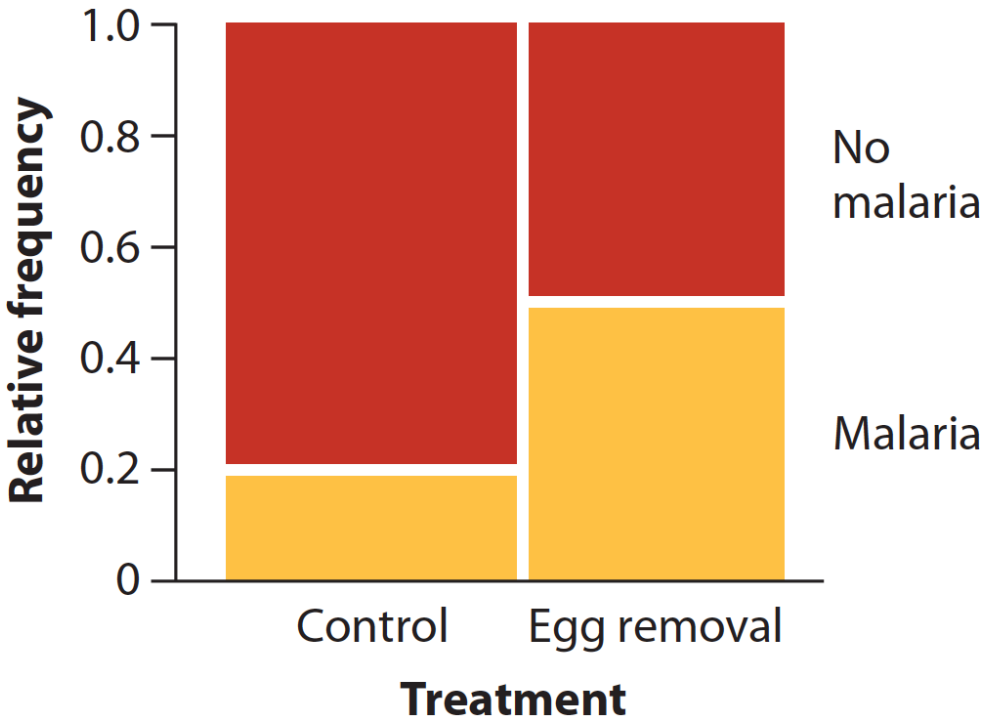
- Test the independence of two or more categorical variables

	Control Group	Egg-Removal Group	Row Total
Malaria	7	15	22
No Malaria	28	15	43
Column Total	36	30	65

Contingency Analysis:

- Associations between categorical variables
- Test the independence of two or more categorical variables

	Control Group	Egg-Removal Group	Row Total
Malaria	7	15	22
No Malaria	28	15	43
Column Total	36	30	65



Reminder: Multiplication Rule

Multiplication rule: $P[A \text{ and } B] = P[A|B]P[B]$

IFF INDEPENDENT, this collapses to: $P[A \text{ and } B] = P[A]P[B]$

χ^2 Contingency Test:

- Tests goodness-of-fit to the data of the null hypothesis of independence of variables
- Two categorical variables but, unlike the Odds Ratio, each variable can have more than 2 categories
- Assumptions:
 - The value of the cell **expected values** should be 5 or more in at least 80% of the cells
 - No cell should have an **expected value** of less than one
- Description of χ^2 Contingency Test:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/>

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

STEP 1: Formulate null hypothesis

Example: Is there a relationship between age at first birth and the development of breast cancer?

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

Step 1:

H_0 : The development of breast cancer is ***independent*** of the age at first birth

H_A : The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence. Assumptions: no cells less than 5 so both assumptions are met.

With independence,

$P[\text{Age at first birth AND breast cancer}] = ?$

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

Step 1:

H₀: The development of breast cancer is ***independent*** of the age at first birth

H_A: The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence

With independence,

$$P[\text{Particular Age at first birth AND breast cancer}] = P[\text{Particular Age at first birth}]P[\text{Breast cancer}]$$

Calculating the expectations under H_0 :

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

$$P[Age < 20 Birth] = \frac{1742}{13465} = 0.13$$

$$P[Cancer] = \frac{3220}{13465} = 0.24$$

$$P[No Cancer] = \frac{10245}{13465} = 0.76$$

If H_0 is true, then:

$$P[< 20 \text{ Age at first birth AND breast cancer}] = 0.13 * 0.24 = 0.031$$

$$\text{Expected count } <20 \text{ and cancer cell} = 0.031 * 13465$$

Calculating the expected **COUNTS** under H_0 :

EXPECTED values Under H_0	<20	20-29	30-34	>=35	Row total
Cancer	416.6 320	2281.9 2217	371.9 463	149.7 220	3220
No Cancer	1325.6 1422	7260.2 7325	1183.2 1092	477 406	10245
Column Total	1742	9542	1555	626	13465

χ^2 Contingency Test

Step 2:

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} = 104.76$$
$$= \frac{(416.6 - 320)^2}{416.6} + \frac{(2281.9 - 2217)^2}{2281.9} + \frac{(371.9 - 463)^2}{371.9} + \frac{(149.7 - 220)^2}{149.7} + \frac{(1325.6 - 1422)^2}{1325.6} + \frac{(7260.2 - 7325)^2}{7260.2} + \frac{(1183.2 - 1092)^2}{1183.2} + \frac{(477 - 406)^2}{477}$$

Step 3:

Degrees of Freedom:

$$\text{dof} = (\text{row} - 1)(\text{column} - 1)$$

For the Birth age/cancer example: **dof = (2-1)(4-1)=3**

Step 4, Conclusion:

$$\chi^2 = 104.76 \gg \chi^2_3 = 7.81$$

We reject the null hypothesis of independence with a significance level of $\alpha = 0.05$ and say that the age of first birth was not independent on whether breast cancer eventually developed.

Is there an influence of the following three SES on preterm delivery rates?

Socio-Economic status	Preterm Birth	Normal Birth
Upper/Upper-middle	25	85
Middle	33	64
Lower/Lower-middle	112	149

- A. Yes, we reject the null hypothesis
- B. No, we fail to reject the null hypothesis
- C. Yes, we fail to reject the null hypothesis
- D. No, we reject the null hypothesis