

# $\chi^2$ Test of Independence or Test of Association

*Bharat Pokharel*

*School of Forest Resources and Environmental Science*

FW3200: Inventory, Monitoring and Data Analysis



## Outline – Chi-Square Test of Independence or Association

---

- What is it?
- Why I do it?
- What are the assumptions
- How I do it?
- Example along with hypothesis



## Why I am doing it?

---

### Statistical Tests

Independent variable (X)	Dependent variable (Y)	
	Continuous (Quantitative)	Categorical (Discrete or qualitative)
Continuous (Quantitative)	Linear Regression	Logistic regression
Categorical (Discrete or qualitative)	T test ANOVA	$\chi^2$ (Chi-Square test)



## $\chi^2$ Test of Independence

---

- Observations are categories, which are recorded in terms of frequency
- We organize the data in a contingency table using frequency under each category
- Question we are asking here is: are these two categories independent or associated or related?
- For example, gender and choice of pets; political affiliation and favor of gun laws in the US, stream velocity and streambed substrate



## Assumptions

---

- There is no assumptions associated with this test as long as the data are recorded in frequencies
- All expected values should be  $>1$  and no more than 20% of expected values should be less than 5.



## How do we conduct this test?

---

1. First sort out your data by two-way contingency table, make sure data are expressed in frequencies (count, e.g use pivot table in excel)
2. These values from 1) are observed values ( $O_i$ ), means observations from sampling, field measurement or experiment
3. Based on observed value calculate expected values for each category ( $E_i$ )



## Expected values ( $E_i$ )

---

1. Statistical independence means joint probability equals product of marginal probabilities
2. Compute marginal probabilities and multiply for joint probability
3. Expected value is the total sample size times the joint probability
4. This is same as multiplying row total by column total, then divide it by the grand total sample size



## Example

---

- We randomly select a group of 50 female and 50 male between the ages of 5 and 50 who visit a pet shop over the last summer and record which animal they choose as their first pet (Sample)

Name	Gender	Pet choice
Mike	M	Turtle
Linda	F	Puppy
Bethany	F	Hamster
..		
..		



## Example

---

- Hypotheses
  - $H_0$ : Variables are independent
  - $H_a$ : Variables are related (dependent)

Gender	Pet Choice			Row Total
	Puppy	Turtle	Hamster	
Female	10	13	27	50
Male	23	16	11	50
Column total	33	29	38	100



## Example

---

Gender	Pet Choice			Row Total
	Puppy	Turtle	Hamster	
Female	10 $(33 \cdot 50) / 100 = 16.5$	13 $(29 \cdot 50) / 100 = 14.5$	27 $(38 \cdot 50) / 100 = 19$	50
Male	23 $(33 \cdot 50) / 100 = 16.5$	16 $(29 \cdot 50) / 100 = 14.5$	11 $(38 \cdot 50) / 100 = 19$	50
Column total	33	29	38	100



## $\chi^2$ Calc

---

$$\chi^2_{calc} = \sum \frac{(O - E)^2}{E}$$

$$df = (\# \text{ of rows} - 1)(\# \text{ of columns} - 1)$$

If  $\chi^2_{calc} > \chi^2_{cri}$  then reject  $H_0$



## $\chi^2$ Calc

---

$$\chi^2 \text{ Calc} = (10 - 16.5)^2 / 16.5 \dots\dots\dots$$

$$\dots\dots\dots (11 - 19)^2 / 19 = \mathbf{12.8}$$

$$\chi^2 \text{ Crit} = \mathbf{5.99}, \text{ at } \alpha = 0.05 \text{ at } (2 - 1)(3 - 1) = 2 \text{ df}$$

Then, we **reject  $H_0$** , gender and pet choice are dependent

### **Conclusion:**

The selection of a first pet is contingent upon, or dependent upon a person's gender at 5 percent significance level

