July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp
with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

For further details contact

Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix health systems

Non-Parametric Tests

Introduction

- Statistical tests can be grouped into 2:
  - Parametric
  - Non-parametric
- Parametric tests are applicable to variables that assume normal distribution in the population of study
- Non-parametric tests are distribution free and are applicable to categorical and ordinal variables.

Types of non-parametric tests

- Chi-squared (χ^2) tests
  - Yate's correction
  - Fisher exact
  - Mantel-Haenszel
  - McNemar
- Kruskal-Wallis *(see lecture on analysis of variance)*
- Logistic regression *(see lecture on logistic regression)*

Uncorrected chi-squared (χ^2) test

- This is the most commonly used non-parametric test.
- The test is usually used to look for association between different categories
- Both independent and dependent variables are categorical.
- The χ^2 test is based on measuring the difference between the observed frequencies and the expected frequencies
- If the null hypothesis is true there should be no difference in the observed and expected frequencies in the groups being compared

Steps in performing the test

1. Decide on the level of significance usually referred to as the P-value i.e. (α-level); usually set at 5% (0.05) or 1% (0.01)
2. Calculate the χ^2 statistic
3. Use a χ^2 table to find the critical value
4. Interpret the result, make a decision

1. Calculate the expected frequency (E) for each cell:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total}}$$

2. For each cell subtract Expected frequency (E) from Observed frequency (O) i.e., O – E

3. Square the result, i.e., $(O - E)^2$ and divide by the expected frequency (E) i.e.,

$$\frac{(O - E)^2}{E}$$

4. Add the result of 3. for all cells to get a value for the χ^2:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_n - E_n)^2}{E_n}$$

Use the χ^2 table

- Determine the degree of freedom (df) which is usually (number of rows – 1)× (number of columns – 1)
- Check the χ^2 value in the χ^2 table using the agreed P-value and the degree of freedom
- Interpret the result by comparing the calculated χ^2 value with the χ^2 value from the χ^2 table.
- If the calculated value is higher than the value from the table, then the P-value is less than what has been set and the observed difference is said to be significant.

Worked example

> Suppose that in a study of the factors affecting the utilization of antenatal clinics you found that 64% of 80 women who lived within 10 km of a clinic came for antenatal care compared to only 47% of 75 women who lived more than 10 km away.

- This suggests that antenatal care (ANC) is used more often by women who live close to the clinics.
- Is this difference in utilization significant?

Observed frequencies

Distance from ANC	Used ANC	Did not use ANC	Total
Less than 10 km	51	29	80
10 Km or more	35	40	75
Total	86	69	155

Expected frequencies

Distance from ANC	Used ANC	Did not use ANC	Total
Less than 10 km	**44.4**	**35.6**	80
10 Km or more	**41.6**	**33.4**	75
Total	86	69	155

Calculate the χ^2 statistic

$$\chi^2 = \frac{(51 - 44.4)^2}{44.4} + \frac{(29 - 35.6)^2}{35.6} + \frac{(35 - 41.6)^2}{41.6} + \frac{(40 - 33.4)^2}{33.4}$$

$$\chi^2 = 0.98 + 1.22 + 1.05 + 1.30 = 4.55$$

Use the χ^2 table

- Determine the degree of freedom (df): (2 rows – 1)×(2 columns – 1) = 1 (df = 1).
- Look up the value of theoretical χ^2 on the table and compare with the calculated χ^2 and identify the χ^2 with the biggest value.
- χ^2 at df = 1 and P = 0.05 is 3.84.

Interpret the result

- Using the table of χ^2, with a df of 1, the calculated χ^2 of 4.55 is larger than 3.84 from the table. This means that the P value is less than 0.05.

> Women living within a distance of 10 km from the clinic use antenatal care significantly more often than the women living more than 10 km away.

2 × 2 contingency table

		Condition		
		Yes	No	Total
Exposure	Yes	a	b	***e***
	No	c	d	***f***
	Total	***g***	***h***	***n***

$$\chi^2 = \frac{n(ad - bc)^2}{efgh}$$

Limitations

- The uncorrected χ^2 test is only valid when all expected values in cells are reasonably large:
- At least 5 for a 2 × table.
- The χ^2 test is not appropriate for quantitative data.
- Sample size should be at least 40

Yate's correction

Suppose we study mortality in Malaria and find the following results:

- Survival CQ vs Quinine: 96.2% vs 92.6%.
- Is this difference significant?

Observed frequencies

	Survived	Died	Total
CQ	75	3	78
Quinine	75	6	81
Total	150	9	159

Expected frequencies

	Survived	Died	Total
CQ	**73.6**	**4.4**	80
Quinine	**76.4**	**4.6**	75
Total	150	9	159

- The uncorrected $\chi^2 = 0.944$ (< 3.84), so $P > 0.05$
- The expected values in the Chloroquine and Quinine died cells are < 5
- The general rule is that the χ^2 test is not valid when an expected value < 5
- However, you can use the Yates correction in a 2×2 table to ***partially*** compensate for low expected values in cell as long as most of the cells have expected values greater than five.

- To make the correction, change $(O - E)^2/E$ for the affected cells to $(O - E + 0.5)^2/E$.
- So e.g. in the Chloroquine-Died cell above, we would use:
  - $(3 - 4.4 + 0.5)^2/4.4 = (1.4 - 0.5)^2/4.4 = 0.184$,
  - Rather than the uncorrected value of $(3 - 4.4)^2/4.4 = 0.445$.

Yate's correction

		Condition		
		Yes	No	Total
Exposure	Yes	a	b	***e***
	No	c	d	***f***
	Total	***g***	***h***	***n***

$$\chi^2 = \frac{(n-1)\left(|ad - bc| - \frac{n}{2}\right)^2}{efgh}$$

Mantel-Haenszel χ^2 test

Dealing with confounding

> Supposing that in survey of Schistosomiasis in two villages ***A*** and ***B*** the research obtained the following results:

	Yes	**No**	**Total**
Village *A*	80	170	250
Village *B*	80	170	250
Total	160	340	500

- Prevalence is the same in both villages (32%)
- $\chi^2 = 0.009$, 1 df, $p > 0.05$

Although the prevalence is the same in both villages the researcher suspects that age may be a confounding variable and he decided to break the table into 2:

- 5 - 19 years of age
- 20 years and above

Children 5-19 years

	Yes	No	Total
Village *A*	37	23	60
Village *B*	73	117	190
Total	110	140	250

$\chi^2 = 9.08$; 1 df; $p < 0.001$

Adults 20 years and above

	Yes	No	Total
Village *A*	43	147	190
Village *B*	7	53	60
Total	50	200	250

$\chi^2 = 2.78$; 1 df; $p > 0.05$

- Breaking the data down according to the confounding variable age (stratification) and applying the Mantel-Haenszel χ^2,
- The prevalence of *schistosomiasis* in children was statistically significantly different from the prevalence in adults ($\chi^2 = 9.08$; df = 1; $P < 0.001$).
- There are more children in village B and schistosomiasis is more common in children in village A than B.

> Age in this case is a confounding variable because it affects the variable of interest (prevalence of schistosomiasis) and the groups being compared (residence in villages A or B)

McNemar chi-squared (χ^2) test

Dealing with paired categorical data

- In dealing with paired observation that are categorical, both the usual χ^2 and the Mantel-Haenszel χ^2 test are not appropriate
- The appropriate test is McNemar's χ^2 test which is used mainly for nominal data to compare proportions of paired observations

- In an outbreak of cholera in a community, a study was conducted to identify the causes of the cholera.
    - For each cholera case, a subject was sought of the same age decade, sex and the same neighborhood, i.e., matching the case with suitable control.
    - Each case and its control are treated as a pair and information obtained on seafood consumption of by each case and control as a pair

Ate sea food?

		Controls		
		+	–	Total
Cases	+	12	30	42
	–	3	31	34
	Total	15	61	76

- 12 pairs of both cases and control ate seafood
- 31 pairs did not eat seafood.
- These 12 + 31 = 43 pairs give us no information about whether eating seafood is a risk factor for getting cholera.
- In 30 pairs (39%), the cases ate seafood but the control did not
- In 3 pairs (4%), cases did not eat seafood but the control did.
- It seems eating seafood is a risk factor for getting cholera.
    - This assumption is only true provided that the McNemar χ^2 test gives a significant result.

$$\chi^2 = \frac{(|r - s| - 1)^2}{r + s}$$

$$\chi^2 = \frac{(|30 - 3| - 1)^2}{30 + 3}$$

- r is the number of +– responses
- s is the number of –+ responses.
- |r – s| is the absolute difference between r and s.
- McNemar's test should be used only when r + s ≥ 20

Fisher's exact test

- Another way of handling small expected values in 2×2 tables is to use the Fisher’s exact test.
- This uses probability theory to calculate the exact probability of obtaining the observed or greater departure from the expected.
- This test is very suitable for sample sizes < 40 where χ^2 test is not recommended.
- The test treats data in a manner similar to χ^2 test.

		Condition		
		Yes	No	Total
Exposure	Yes	a	b	***e***
	No	c	d	***f***
	Total	***g***	***h***	***n***

when a<ge/n

$$\sum_{k=\max(0,g-f)}^{a} \frac{\binom{e}{k}\binom{f}{g-k}}{\binom{n}{g}}$$

when a>ge/n

$$\sum_{k=a}^{\min(e,g)} \frac{\binom{e}{k}\binom{f}{g-k}}{\binom{n}{g}}$$

Conclusion

- Nowadays the use of statistical tests for significance testing has been made simple by the use of computers.
- All one needs to know these days is how give the appropriate command and the computer does the analysis.
- However one must know what to look for to be able to make the right interpretation.