Chi-Squared Tests

Collin Nolte

March 01, 2022

Review

In the previous class, we considered statistical tests for determining differences between groups on a continuous variable

Frequently, however, our data will be categorical in nature. These are typically arranged in tables of varying sizes and complexity

Our interest today will consider of subset of this problem, represented with a 2×2 table

2×2 table

Population	Variable		Total
	$+$	$-$	
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

Sampling Methods

There are primarily three ways in which we may sample data that can be represented as a 2×2 table

1. We decide a priori to sample and measure (count) two characteristics:
   - The only fixed value is N, the total size of the sample
   - These situations are associated with *tests of independence*
   - Typically an observational study
   - E.g., draw a random sample of N people and classify them according to sex and phenotype status (binary)

Sampling Methods

There are primarily three ways in which we may sample data that can be represented as a 2×2 table

2. Draw independent samples from two populations and then look for presence or absence of phenotype:

- Here, N is fixed, as well as one set of margin totals
- These situations are associated with *tests of homogeneity*
- Typically an observational study
- E.g., draw a random sample of n_1 males and n_2 females and classify them according phenotype status (binary)

Sampling Methods

There are primarily three ways in which we may sample data that can be represented as a 2×2 table

3. Draw a set of N subjects and assign them randomly to two groups (i.e., case and control). Treatment is applied to one and not the other:

- Here, N is fixed, as well as one set of margin totals
- These situations are associated with *tests of homogeneity*
- This would be a prospective study (i.e., clinical trial)
- Randomize group of N individuals into two treatment groups, carry out the treatment, and then identify phenotype status at end of study

Independence vs Homogeneity

Population	Variable		Total
	+	−	
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

Independence: Probability of being in A *and* being positive is equal to probability of being in A *times* probability of being positive

$$\frac{a}{N} = \frac{a+b}{N} \times \frac{a+c}{N}$$

Homogeneity: Probability of being positive in A is equal to probability of being positive in B

$$\frac{a}{a+b} = \frac{c}{c+d}$$

χ^2 Test Statistic

Although the theoretical development of each differs, the significance test is the same for each

$$\chi^2 = \frac{N \times (ad - bc)^2}{(a+c)(b+d)(c+d)(a+b)} = \sum \frac{(Observed - Expected)^2}{Expected}$$

The statistic is derived by summing the squared difference between the observed counts in each cell and the expected counts, under the null hypothesis

This right hand side of the equality is especially useful for tables larger than 2×2, where the χ^2 statistic is computed in exactly the same way

χ^2 Test Statistic

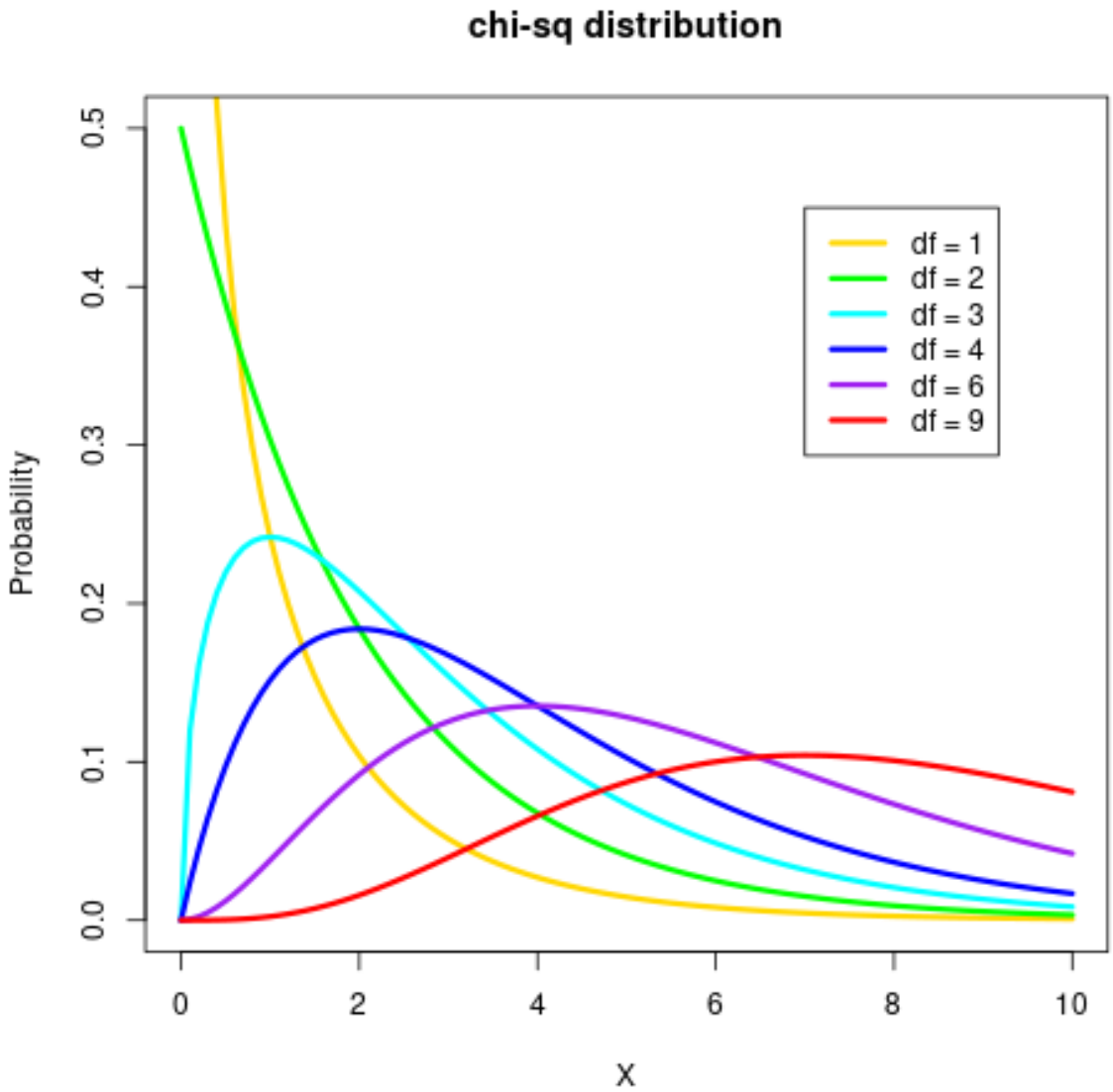
It's important to note here that the χ^2 distribution is a *continuous* distribution, and we are using it to *approximate* the discrete sampling that actually occurs

The parameter associated with the χ^2 distribution is *degrees of freedom*. In general, we have

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

For 2×2 tables, the degrees of freedom will always be $df = 1$

χ^2 Distribution



Example: Test of Independence

Counties with	Stork Count		Total
	Few Storks	Many Storks	
Low Birthrate	O = 42 E = 44.72	O = 10 E = 7.28	52
High Birthrate	O = 130 E = 127.28	O = 18 E = 20.72	148
Total	172	28	200

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 1.60$$

$$p = Pr\{\chi_1^2 \geq 1.60\} \approx .206$$

Example: Test of Independence

Counties with	Stork Count		Total
	Few Storks	Many Storks	
Low Birthrate	O = 42 E = 44.72	O = 10 E = 7.28	52
High Birthrate	O = 130 E = 127.28	O = 18 E = 20.72	148
Total	172	28	200

As 52 counties have low birthrate out of 200 total, we note that low birthrate counties make up $52/200 = 26\%$ of total counties

172 counties have a low number of storks. If stork count and birthrate were independent, we would then expect that 25% of low stork counties would have low birthrate. $172 \times 0.26 = 44.72$. Similarly for high stork counties, we have $28 \times 0.26 = 7.28$

Example: Test of Homogeneity

Treatment Status	Side Effects		Total
	Yes	No	
Control	O = 10 E = 20	O = 40 E = 30	50
Treatment	O = 30 E = 20	O = 20 E = 30	50
Total	40	60	100

Here, we consider a study in which a population of $N = 100$ was sampled, and subjects were randomly assigned to either a control or treatment group. Researchers are interested in knowing if any negative side effects were reported over the course of the study.

$$\chi^2 = \frac{100(10 \cdot 20 - 40 \cdot 30)^2}{50 \cdot 40 \cdot 50 \cdot 60} = 16.67$$

$$p = Pr\{\chi_1^2 \geq 16.67\} \approx 0.00004$$

Sample Size and χ^2

It is of exceptional importance to always recall that this is a binary test: the evidence gathered and our conclusion can only be stated in terms of independence/homogeneity. In other words, *the p-value makes absolutely no claim about the magnitude of association between two groups*

$$\chi^2 = \frac{N \times (ad - bc)^2}{(a+c)(b+d)(c+d)(a+b)}$$

Case 1:

Treatment Status	Side Effects		Total
	Yes	No	
Control	10	40	50
Treatment	30	20	50
Total	40	60	100

$$\chi^2 = \frac{100(10 \cdot 20 - 40 \cdot 30)^2}{40 \cdot 60 \cdot 50 \cdot 50} = 16.67$$

Case 2:

Treatment Status	Side Effects		Total
	Yes	No	
Control	100	400	500
Treatment	300	200	500
Total	400	600	1000

$$\chi^2 = \frac{1000(100 \cdot 200 - 400 \cdot 300)^2}{400 \cdot 600 \cdot 500 \cdot 500} = 166.67$$

The χ^2 Approximation

Another limitation of the χ^2 approximation can be observed if any of the cells have an *expected count* of less than five.

While there are "corrections" to attempt to adjust for this, their use is controversial and we can do one better. While often computationally expensive for larger tables, in the 2×2 case we can use Fisher's Exact test

Unlike the χ^2 approximation, Fisher's test is based on the hypergeometric distribution for discrete data

Hypergeometric distribution

The motivation of the hypergeometric distribution is as follows: Suppose we have N marbles in an urn, and exactly K of them are black. If I draw $n \leq N$ marbles from the urn without replacing any, what is the probability that I draw exactly $k \leq K$ black ones?

Fisher Exact Test

Population	Variable		Total
	+	−	
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

By analogy, we have a population of N observations, with $(a+b)$ draws from our population with $(a+c)$ positive phenotypes and $(b+d)$ negative phenotypes. What is the probability of having drawn the value a?

From the previous slide, this maps to $N = N$, $K = (a+c)$, $n = (a+b)$ and $k = a$

Fisher Exact Test

The reported p value then gives us a probability of having obtained the arrangement in the observed table, assuming that the row and column totals remain the same. We have

$$p = \frac{(a+c)!\ (b+d)!\ (a+b)!\ (c+d)!}{a!\ b!\ c!\ d!\ N!}$$

This tricky bit here involves finding all of the other tables that would be considered "more extreme", the probabilities of which we would add together

Observed: $p = 0.000191$

Population	Pregnancy Deliveries		Total
	Term	Premature	
Drug	8	10	18
Placebo	24	1	25
Total	32	11	43

More Extreme: $p = 0.0000055$

Population	Pregnancy Deliveries		Total
	Term	Premature	
Drug	7	11	18
Placebo	25	0	25
Total	32	11	43

More Extreme: ...

R

```
> fisher.test(pregnancy)

        Fisher's Exact Test for Count Data

data:  pregnancy
p-value = 0.0002
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.00074489 0.32222602
sample estimates:
odds ratio
  0.036609
```

R

```
> chisq.test(pregnancy)

        Pearson's Chi-squared test with Yates'
        continuity correction

data:  pregnancy
X-squared = 12, df = 1, p-value = 0.00052

Warning message:
In chisq.test(pregnancy) : Chi-squared approximation
may be incorrect
```

McNemar's Test for Matched Pairs

Suppose now we consider an example in which we collect a participants for a focus group and ask whether or not they favor a proposed bond issue.

Following a media ad campaign, the same individuals are interviewed and once again asked their opinion on the issue

Before	After		Total
	+	-	
+	50	9	59
-	21	20	41
Total	71	29	100

Clearly, the results before and after the media campaign are not independent

Matched Pairs

Before	After +	After -	Total
+	a	c	a+b
-	b	d	b+c
Total	a+c	b+d	2N

What we are interested in here are the *shifts* that occurred between the two measurements. Such pairs are called *discordant*

If the media campaign had no impact on public opinion, we should expect that individuals are equally likely to change their mind, regardless of what they had thought before

This is known as a *test of symmetry*. Under our null hypothesis, we have $H_0 : \{\text{Probability of changing mind}\} = p = 1/2$

Tests

As before, we have two tests we can use: the first utilizes a χ^2 approximation, the second computes an exact p-value

McNemar's Test:

- χ^2_1 approximation
- Requires $(b + c) > 25$

$$\chi^2_1 = \frac{[|b - c| - 1]^2]}{b + c}$$

Exact Binomial Test:

- Two outcomes, either $\{+ \to -\}$ or $\{- \to +\}$
- Assumes $H_0 : p = 0.5$
- Works regardless of sample size

Before	After +	After -	Total
+	50	9	59
-	21	20	41
Total	71	29	100

```
> mcnemar.test(m)

        McNemar's Chi-squared test with
        continuity correction

data:  m
McNemar's chi-squared = 4.03, df = 1, p-value = 0.045
```

Before	After +	After -	Total
+	50	9	59
-	21	20	41
Total	71	29	100

```
> binom.test(x = 21, n = 9 + 21)

        Exact binomial test

data:  21 and 9 + 21
number of successes = 21, number of trials = 30, p-value = 0.043
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.50604 0.85265
sample estimates:
probability of success
                   0.7
```

Application in Genetics

An implementation of McNemar's test in human genetics is the Transmission Disequilibrium Test (TDT)

It is an AFBAC (**A**ffected **F**amily-**BA**sed **C**ontrol) method used in family-based association studies

Note: FBAT = Family Based Association Test

TDT Example

Basic Idea: We will use non-transmitted parental alleles in a triad composed of an *affected child* and the two parents. Red indicates the transmitted allele

Parent 1 Parent 2

$M1M1 \times M2M2$

$\updownarrow$

$M1M2$

Child

Parent 1 Parent 2

$M1M2 \times M1M2$

$\updownarrow$

$M1M2$

Child

Allele Transmission Table

Recall here that we are only considering *affected children*

Transmitted	Non-transmitted		Total
	M1	M2	
M1	a	b	
M2	c	d	
Total			2N

Homozygous parents (either $M1M1$ or $M2M2$) will always transmit and not transmit the same allele. Heterozygous parents, on the other hand, will always trasmit one and not the other.

If there is no association between the expressed phenotype and the transmitted allele, we would expect an equal proportion of $M1$ genes transmitted as the expense of $M2$ as we would $M2$ genes being trasnmitted at the expense of $M1$ from heterozygous parents

TDT

As before, both McNemar's test and the exact binomial are relevant here

This test is apprporiate regardless of penetrance values and ascertainment procedure and is *not* affected by population stratification

The TDT assumes no segregation distortion at the marker locus (see Parsian et al, 1991 for alternatives). Further, multiple extensions have been created to encompass multiallelic markers, quantitative traits, and the inclusion of additional family members (multiple affected siblings)

References

Deb Dawson course notes

Falk CT, Rubinstein P (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51 (Pt 3):227-233.

Parsian A, Todd RD, Devor EJ, O'Malley KL, Suarez BK, Reich T, Cloninger CR (1991). Alcoholism and alleles of the human D2 dopamine receptor locus. Studies of association and linkage. Arch Gen Psychiatry. 48:655–663.

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-516.