STAT 120 C

Introduction to Probability and Statistics III

Dustin Pluta
2019/05/29

Categorical Data Analysis

In categorical data analysis, we consider observations that belong to sets of categories.

Examples

- Are carriers of a particular gene more susceptible to cancer?
- Is heart attack incidence associated with blood type?
- Is gender associated with likelihood of promotion?
- Do STEM degrees or humanities degrees have lower unemployment?

Categorical Data Analysis

- **Nominal** variables have categorical values
- For example, *Blood Type* takes values over 8 categories (A+, A-, B+, B-, O+, O-, AB+, AB-).
- Each person's measured blood type will belong to exactly one of these categories.
- A sample of 100 individuals may have blood type distributed as:

Blood Type	A+	A-	B+	B-	O+	O-	AB+	AB-
Count	34	40	7	3	8	7	1	0

Review of Multinomial Distribution

- Consider a random vector of count data $X = (N_1, N_2, \ldots, N_c)$, where each N_i is a count of elements in category i.

- X follows a multinomial distribution if it has probability mass function

$$p(N_1 = n_1, N_2 = n_2, \cdots, N_c = n_c) = \left(\frac{n!}{n_1! n_2! \cdots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c},$$

where

$$\sum_{i=1}^{c} n_i = n$$

$$\sum_{i=1}^{c} \pi_i = 1$$

- We write $X \sim Multi(n, (\pi_1, \ldots, \pi_c))$.

Review of Multinomial Distribution

Properties

$$\mathbb{E}(N_i) = n\pi_i$$

$$\text{Var}(N_i) = n\pi_i(1 - \pi_i)$$

$$\text{Cov}(N_i, N_j) = -n\pi_i\pi_j, i \neq j$$

Marginal Distribution

$$N_i \sim Binom(n, \pi_i)$$

Conditional Distribution

$$(N_1, \cdots, N_{c-1})|(N_c = n_c) \sim Multi\left(n - n_c, \frac{pi_1}{1 - \pi_c}, \ldots, \frac{\pi_{c-1}}{1 - \pi_c}\right)$$

Note that when $c = 2$, the multinomial distribution reduces to the binomial distribution.

Review of Multinomial Distribution

Example

The distribution of the blood type data may follow a multinomial:

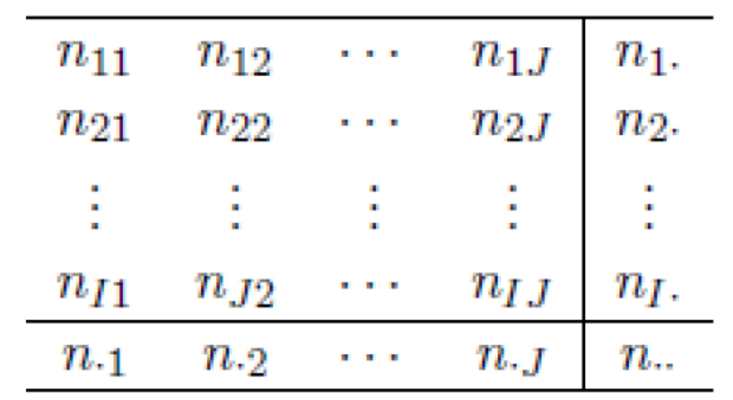
$$X \sim Multi(100, (0.374, 0.357, 0.085, 0.034, 0.066, 0.063, 0.015, 0.006))$$

The count data from the table is a realization of this random variable from a random sample of 100 people.

Blood Type	A+	A-	B+	B-	O+	O-	AB+	AB-
Count	34	40	7	3	8	7	1	0

Pearson's Chi-squared Test

Consider a two-way contingency table



n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{I1}	n_{J2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

where

- n_{ij} is the observed count in row i and column j
- $n_{i\cdot} = \sum_{j=1}^{J}$ is the total number of observations in row i
- $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$ is the total number of observations in column j
- $n_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ is the total number of observations

Pearson's Chi-squared Test

- If we have two factors, the Chi-squared test can be used to determine:
  - Are the two factors independent?
  - Are subpopulations homogeneous (i.e. equally distributed)?
- Which question we are answering depends on the data we have and the goal of the analysis.

Pearson's Chi-squared Test

The test statistic is

$$T = \sum_{i=1}^{c} \frac{(Obs_i - Exp_i)^2}{Exp_i},$$

where

- c is the total number of cells (e.g. $c = IJ$ for an $I \times J$ table),
- Obs_i is the observed count for cell i,
- Exp_i is the expected count for cell i under some specific null hypothesis.

- The value of Exp_i can be calculated from the MLE of the parameters under the null hypothesis.

Pearson's Chi-squared Test

Idea of the test

- The Central Limit Theorem tells us that sums of random variables will be approximately normally distributed.
- Quadratic forms of normal random variables will follow a χ^2 distribution.
- Combining these properties, we can derive the (approximate) null distribution for the test statistic

Pearson's Chi-squared Test

Theoretical Justification

- Consider a multinomial sample $(n_1, n_2, \ldots, n_c)$ of size n.
- The marginal distribution of n_i is $Binom(n, \pi_i)$.
- For large n, the CLT tells us that

$$\hat{\pi} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_{c-1}}{n}\right)^T$$

has an approximate multivariate normal distribution.

Pearson's Chi-squared Test

Theoretical Justification

- Let Σ_0 be the null covariance matrix of $\sqrt{n}\hat{\pi}$, and let

$$\pi_0 = (\pi_{10}, \pi_{20}, \ldots, \pi_{c-1,0})^T$$

be the expectation of π under the null hypothesis.

- Then, by the CLT

$$\sqrt{n}(\hat{\pi} - \pi_0) \to \mathcal{N}(0, \Sigma_0)$$

- Therefore, by results on the distribution of quadratic forms

$$n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1} (\hat{\pi} - \pi_0) \to \chi^2_{c-1}$$

Pearson's Chi-squared Test

Theoretical Justification

- The covariance matrix Σ of $\sqrt{n}\hat{\pi}$ has elements

$$\Sigma = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_{c-1} \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_{c-1} \\ \vdots & \vdots & \vdots & \vdots \\ -\pi_1\pi_{c-1} & \dots & \dots & \pi_{c-1}(1-\pi_{c-1}) \end{pmatrix}$$

- Under a null hypothesis giving values $\pi_0 = (\pi_{10}, \dots, \pi_{c-1,0})^T$, we plug in these values into Σ to obtain Σ_0.

- It can be shown that the Pearson Chi-squared test statistic is equal to

$$T = n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1} (\hat{\pi} - \pi_0).$$

- Thus, the null distribution is $T \overset{H_0}{\sim} \chi^2_{c-1}$.

Pearson's Chi-squared Test for Independence

- Suppose we wish to test whether the factors are independent in a two-way contingency table

- The null hypothesis is

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j},$$

by the definition of independence

- The likelihood for the multinomial can be written

$$\mathcal{L}(\pi_{11}, \ldots, \pi_{IJ}) = \left(\frac{n_{\cdot\cdot}!}{n_{11}! \cdots n_{IJ}!} \right) \pi_{11}^{n_{11}} \pi_{12}^{n_{21}} \cdots \pi_{IJ}^{n_{IJ}}$$

Pearson's Chi-squared Test for Independence

- Under the null hypothesis, the likelihood can be written

$$\mathcal{L}_0 \propto \prod_{i=1}^{I} \prod_{j=1}^{J} [\pi_{i\cdot}\pi_{j\cdot}]^{n_{ij}}.$$

- The log-likelihood is

$$\begin{aligned}
\ell_0 &= \log(\mathcal{L}_0) = \sum_{i=1}^{I} \sum_{j=1}^{J} [n_{ij} \log(\pi_{i\cdot}\pi_{\cdot j})] + const \\
&= \sum_{i} \sum_{j} n_{ij} \log(\pi_{i\cdot}) + \sum_{i} \sum_{j} n_{ij} \log(\pi_{\cdot j}) + const \\
&= \sum_{i} n_{i\cdot} \log(\pi_{i\cdot}) + \sum_{j} n_{\cdot j} \log(\pi_{\cdot j}) + const
\end{aligned}$$

Pearson's Chi-squared Test for Independence

- It can be shown that the MLEs are

$$\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}, \quad \text{for } i = 1, \cdots, I$$

$$\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}, \quad \text{for } j = 1, \cdots, J.$$

Pearson's Chi-squared Test for Independence

- Thus,

$$Exp_{ij} = n_{..}\hat{\pi}_{ij} = n_{..}\hat{\pi}_{i.}\hat{\pi}\cdot j = \frac{n_{i.}n_{.j}}{n_{..}}$$

- The resulting χ^2 test statistic can be written

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}.$$

- This statistic has reference distribution $T \overset{H_0}{\sim} \chi^2_{(I-1)(J-1)}$ under the null hypothesis of independence.

Pearson's Chi-squared Test for Independence

Example

200 students were surveyed on their preference between two political candidates *A* and *B*. The following table shows the responses by major subject area.

Observed Counts

	Bio.	**Eng.**	**Soc. Sci.**	**Other**	**Totals**
A	24	24	17	27	92
B	23	14	8	19	64
Undecided	12	10	13	9	44
Totals	59	48	38	55	200

Pearson's Chi-squared Test for Independence

Observed Counts

	Bio.	Eng.	Soc. Sci.	Other	Totals
A	24	24	17	27	92
B	23	14	8	19	64
Undecided	12	10	13	9	44
Totals	59	48	38	55	200

Expected Counts under null hypothesis of independence

	Bio.	Eng.	Soc. Sci.	Other	Totals
A	27.14	22.08	17.48	25.3	92
B	18.88	15.36	12.16	17.60	64
Undecided	12.98	10.56	8.36	12.10	44
Totals	59	48	38	55	200

Pearson's Chi-squared Test for Independence

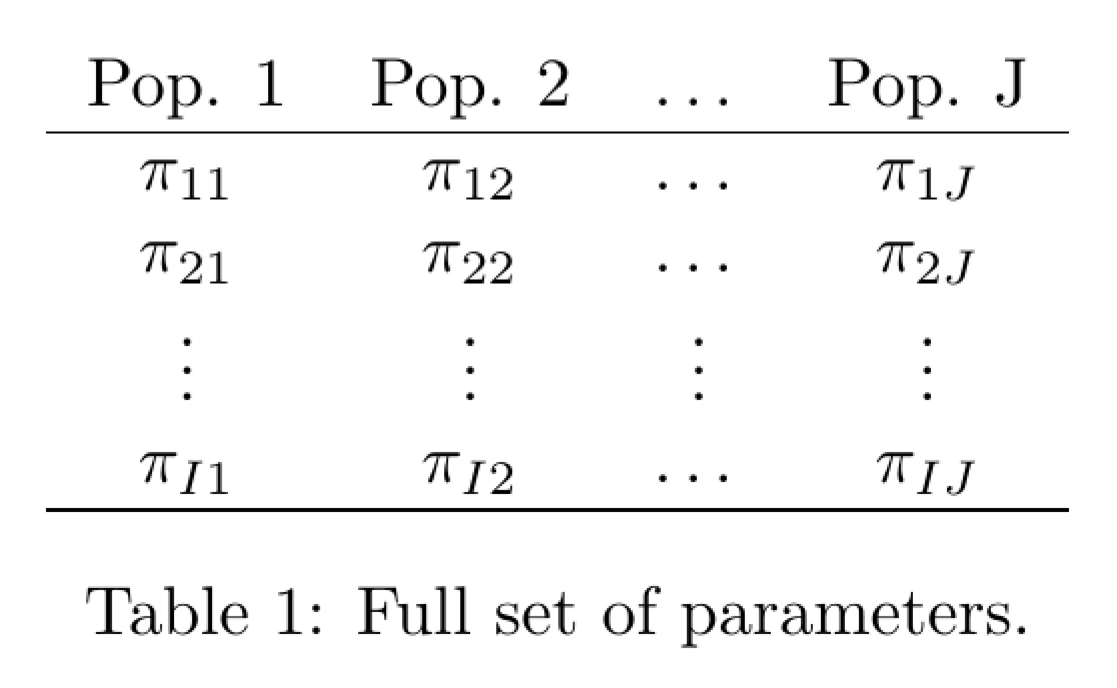
- The statistic is $T = \sum \frac{(Obs - Exp)^2}{Exp} = 6.68$
- Since $I = 3, J = 4$, the null distribution is χ^2_6, where $6 = (I-1)(J-1)$
- The upper 5% tail of χ^2_6 has cutoff 12.59.
- We conclude then that there is insufficient evidence to reject the hypothesis that candidate preference is independent of major type, at significance level 0.05.
- That is, we conclude that candidate preference is associated with major type, at significance level 0.05.

Pearson's Chi-squared Test for Homogeneity

- Suppose now that instead of a simple random sample, the data results from a stratified sample.
- In this setting, we can instead test for *homogeneity* across the subgroups.
- For instance, suppose that in the above example, the data were obtained by sampling 59 students from Bio, 48 students from Soc Sci, etc. (rather than just sampling 200 students randomly).
- The test for homogeneity determines if candidate preferences have the same distribution within each major.

Pearson's Chi-squared Test for Homogeneity

- Let π_{ij} be the probability that an observation in the jth population is in category i.
- Organized as a table, the parameters are:



Pop. 1	Pop. 2	$\ldots$	Pop. J
π_{11}	π_{12}	$\cdots$	π_{1J}
π_{21}	π_{22}	$\cdots$	π_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
π_{I1}	π_{I2}	$\cdots$	π_{IJ}

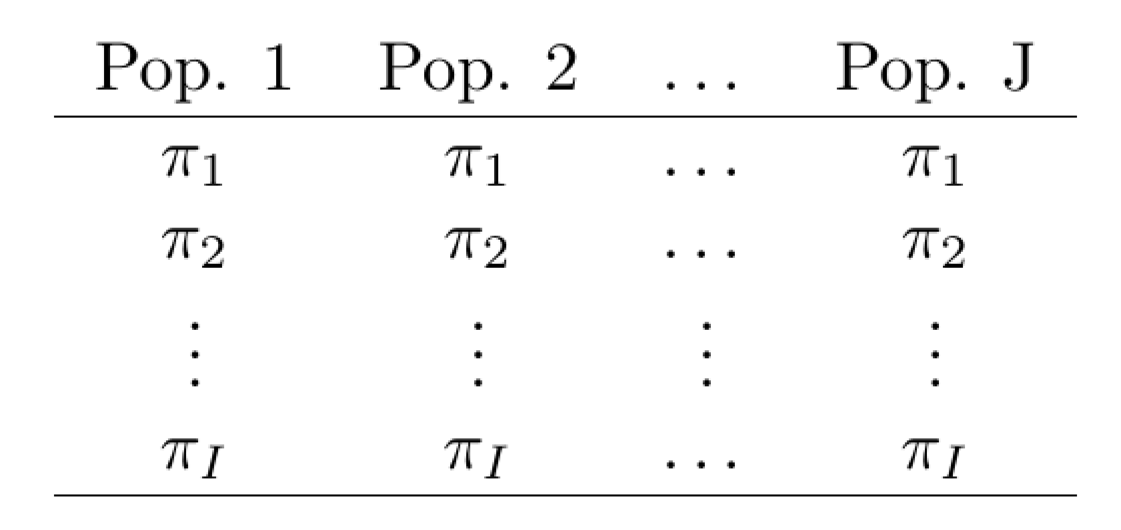
Table 1: Full set of parameters.

Pearson's Chi-squared Test for Homogeneity

- The null hypothesis for the test of homogeneity is

$$H_0 : \pi_{ij} = \pi_i, \quad i = 1, \ldots, I; j = 1, \ldots, J$$

- The parameters under H_0 reduce to



Pop. 1	Pop. 2	$\ldots$	Pop. J
π_1	π_1	$\ldots$	π_1
π_2	π_2	$\ldots$	π_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
π_I	π_I	$\ldots$	π_I

Table 2: Parameters under H_0.

Pearson's Chi-squared Test for Homogeneity

- To test the homogeneity null hypothesis H_0, we can use the Likelihood Ratio Test.
- Using the constraint $\sum_{i=1}^{I} \pi_i = 1$, we can write the null likelihood as

$$\mathcal{L}_0(\pi_1, \ldots, \pi_I) \propto \pi_1^{n_1} \pi_2^{n_2} \ldots (1 - \sum_{i=1}^{I-1} \pi_i)^{n_I}.$$

- The log-likelihood is

$$\ell_0(\pi_1, \ldots, \pi_I) = n_1 \log(\pi_1) + \cdots + n_{I-1} \log(\pi_{I-1}) + n_I \log(1 - \sum_{i=1}^{I-1} \pi_i)$$

Pearson's Chi-squared Test for Homogeneity

- We can find the MLEs through the derivative:

$$\frac{\partial \ell_0}{\partial \pi_i} = \frac{n_{i\cdot}}{\pi_i} - \frac{n_{I\cdot}}{1 - \sum_{i=1}^{I-1} \pi_i} = \frac{n_{i\cdot}}{\pi_i} - \frac{n_{I\cdot}}{\pi_I} = 0$$

$$\hat{\pi}_i = \frac{\hat{\pi}_I n_{i\cdot}}{n_{I\cdot}}$$

- Plugging these estimates into the constraint $\sum \pi_i = 1$,

$$\hat{\pi}_i = \frac{n_{i\cdot}}{n_{\cdot\cdot}}, \quad i = 1, \ldots, I$$

- The expected number of observations in each cell under the null hypothesis is

$$Exp_{ij} = n_{\cdot j}\hat{\pi}_i = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

Pearson's Chi-squared Test for Homogeneity

- The test statistic is then

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}},$$

which has approximate null distribution

$$T \overset{H_0}{\dot{\sim}} \chi^2_{(I-1)(J-1)}.$$

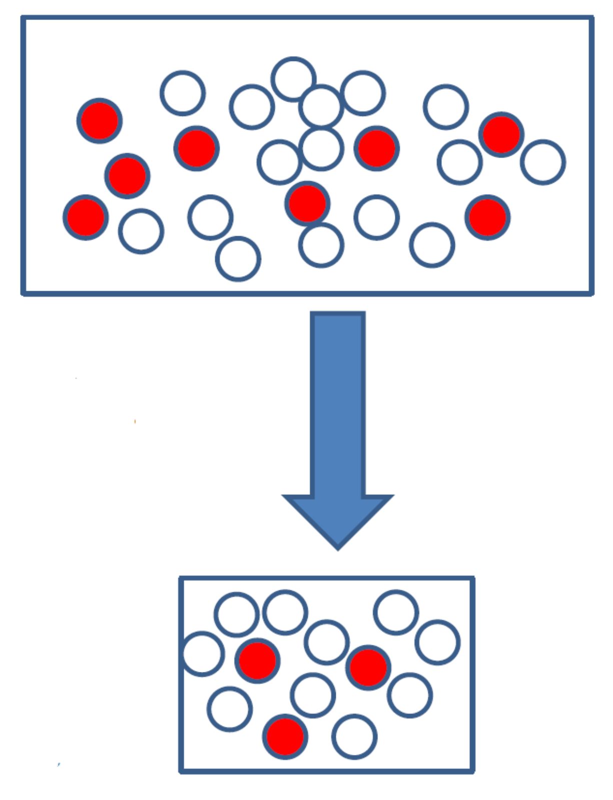
- The degrees of freedom are calculated as the difference of the number of parameters in the full model vs the reduced model:
  - **Full Model**: $(I-1)J$ total free parameters.
  - **Reduced Model**: $(I-1)$ free parameters.
  - Difference $= (I-1)J - (I-1) = (I-1)(J-1)$

Pearson's Chi-squared Tests

- **NOTE** Pearson's tests are based on asymptotic properties, and thus are only appropriate when the sample size is moderately large *in each cell*.
- Often times in tabular data, some cells are sparse, and the results of Pearson's chi-squared test may be invalid
- The usual rule of thumb is to only apply Pearson's chi-squared test when there are at at least 5 observations per cell.
- When we have fewer observations, it is usually better to use an alternative, such as *Fisher's Exact Test*.

Review of Hypergeometric Distribution

- Consider a box containing n balls total, with r red balls, and $n - r$ white balls.
- Suppose we randomly select m balls from the box without replacement.
- The number of red balls selected will follow a hypergeometric distribution.

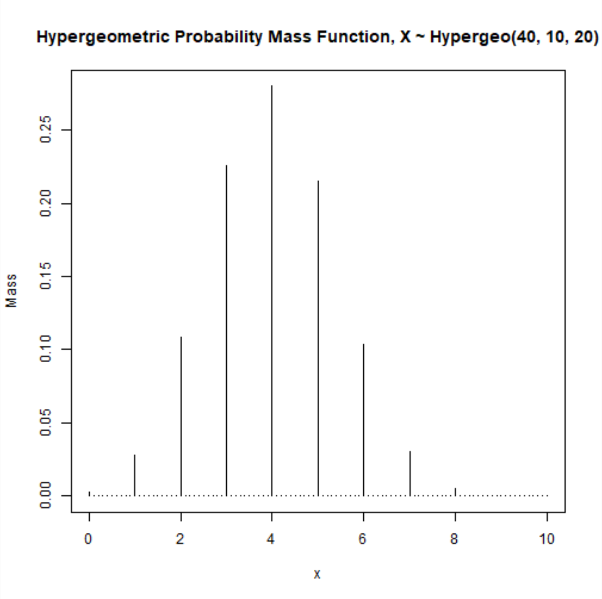


Review of Hypergeometric Distribution

A random variable $X \sim Hypergeo(n, m, r)$ following a hypergeometric distribution has probability mass function

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$$

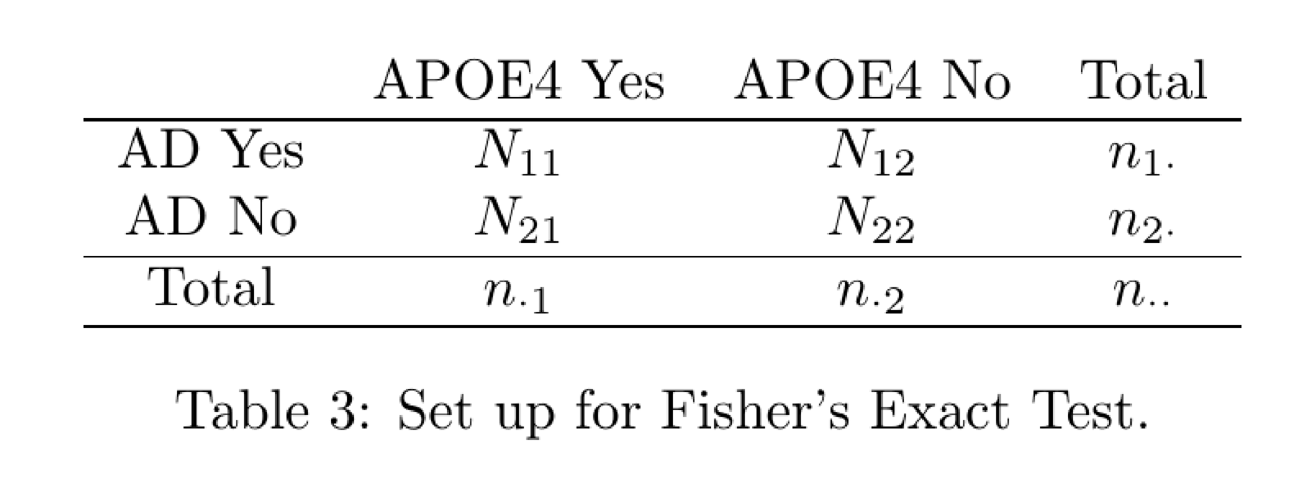
The mean is $\mathbb{E}(X) = m \cdot \frac{r}{n}$.



Fisher's Exact Test

- Suppose we wish to determine whether carriers of the APOE4 gene are more likely to have Alzheimer's Disease by age 70.

- We can consider a sample of data arranged as a 2×2 table.



	APOE4 Yes	APOE4 No	Total
AD Yes	N_{11}	N_{12}	$n_{1\cdot}$
AD No	N_{21}	N_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Table 3: Set up for Fisher's Exact Test.

Fisher's Exact Test

Asummptions

- Independent observations
- Fixed marginal counts $(n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2}, n_{\cdot\cdot})$
- Each observation belongs to exactly one cell in the table

Fisher's Exact Test

Null Hypothesis

- H_0 : There is no association between membership in the row categories and membership in the column categories.
- In our example:
- H_0 : There is no association between carrying the APOE4 gene and occurrence of AD.

Fisher's Exact Test

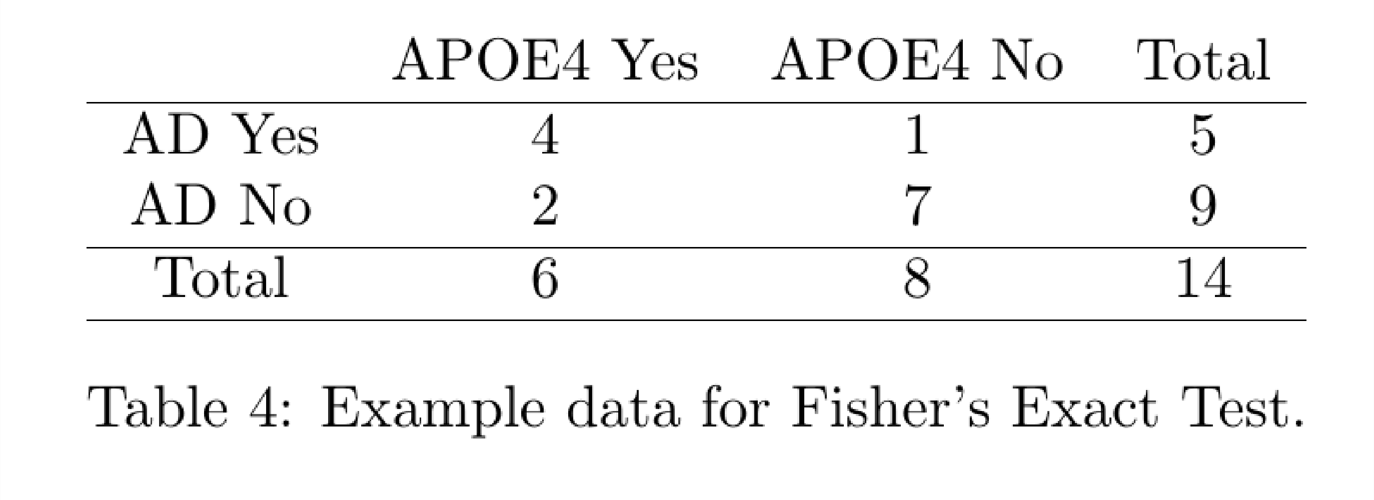
Test Statistic

- There are four random variables in the 2×2 table.
- Because the marginal counts are fixed, if we know one of the cell values, the remaining three cells are fixed.
- Consequently, we can use any of the cell values from observed data as our test statistic, e.g. N_{11}.
- To find the rejection region, we need the distribution of N_{11} under the null hypothesis.
- Under the null hypothesis of no association, N_{11} will follow a hypergeometric distribution.

$$P(N_{11} = n_{11}) \overset{H_0}{=} \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{\cdot 1} - n_{11}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}} = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

Fisher's Exact Test

- Suppose we observe the following data



	APOE4 Yes	APOE4 No	Total
AD Yes	4	1	5
AD No	2	7	9
Total	6	8	14

Table 4: Example data for Fisher's Exact Test.

Fisher's Exact Test

- Under H_0 of no association, the N_{11} cell follows a hypergeometric distribution with $n = 14, m = 6, r = 5$.

- We can compute the P-value for Fisher's Exact test in `R`:

```r
dat <- matrix(c(4, 2, 1, 7), 2, 2)
fisher.test(dat)
```

```
## 
##      Fisher's Exact Test for Count Data
## 
## data:  dat
## p-value = 0.09091
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.6418261 779.1595463
## sample estimates:
## odds ratio 
##   10.98111
```

Fisher's Exact Test

- The P-value is 0.0909, thus we do not have sufficient evidence to reject the null hypothesis at the 0.05 level of significance.

- That is, we do not have sufficient evidence to conclude that there is a significant association between APOE4 and occurrence of AD, at the 0.05 level of significance.