

Relationships between variables

Categorical data

Multivariate statistics

- * We have observations (X_i, Y_i) . Looking for the relationship (if any).
- * Generally, X_i is “independent”, we’re looking to see how Y depends on X .
- * Also call X_i the “predictor” and Y_i the “response”
- * May be categorical or quantitative

Categorical predictor & response: χ^2 test for independence

Example: Handedness by sex, US, aged 25-34.

	Men	Women
Right-handed	934	1070
Left-handed	113	92
Ambidextrous	20	8
Total	1067	1170

source: NHANES 1976-80. Cited Freedman, Pisani, Purves p.537

NHANES* survey: Treat it as a simple random sample of Americans aged 25-34. Sample is small relative to whole population (so samples may be treated as independent)

	Men	Women
Right-handed	934	1070
Left-handed	113	92
Ambidextrous	20	8
Total	1067	1170

* National Health and Nutrition Examination Survey

NHANES survey: Treat it as a simple random sample of Americans aged 25-34

If we think of sex as the independent variable, it makes sense to look at percentages within a column.

Percent	Men	Women
Right-handed	87.5%	91.5%
Left-handed	10.6%	7.9%
Ambidextrous	1.9%	0.7%

Are women more likely to be right-handed?

Null hypothesis: Sex and handedness are independent.

Need: test statistic whose distribution we can compute (approximately) under the null, and whose extreme values correspond to failure of the null.

Percent	Men	Women	Marginal
Right-handed	41.7%	47.8%	89.5%
Left-handed	5.1%	4.1%	9.2%
Ambidextrous	0.9%	0.3%	1.2%
Marginal	47.7%	52.3%	

Null hypothesis: Sex and handedness are independent.

Need: test statistic whose distribution we can compute (approximately) under the null, and whose extreme values correspond to failure of the null.

Idea: Under the null, the probability of being female and left-handed should be the product of the probabilities of being female and the probability of being left-handed.

Percent	Men	Women	Marginal
Right-handed	41.7%	47.8%	89.5%
Left-handed	5.1%	4.1%	9.2%
Ambidextrous	0.9%	0.3%	1.2%
Marginal	47.7%	52.3%	

Null hypothesis: Sex and handedness are independent.

Need: test statistic whose distribution we can compute (approximately) under the null, and whose extreme values correspond to failure of the null.

Idea: Under the null, the probability of being female and left-handed should be the product of the probabilities of being female and the probability of being left-handed.

$$H_0 : \theta_{ij} = \theta_{i\cdot} \cdot \theta_{\cdot j} \quad \theta_{ij} = P(\text{box } ij), \theta_{i\cdot} = P(\text{row } i), \theta_{\cdot j} = P(\text{column } j)$$

Percent	Men	Women	Marginal
Right-handed	41.7%	47.8%	89.5%
Left-handed	5.1%	4.1%	9.2%
Ambidextrous	0.9%	0.3%	1.2%
Marginal	47.7%	52.3%	

Null hypothesis: Sex and handedness are independent.

Need: test statistic whose distribution we can compute (approximately) under the null, and whose extreme values correspond to failure of the null.

Idea: Under the null, the probability of being female and left-handed should be the product of the probabilities of being female and the probability of being left-handed.

	M	W	Mar
RH	41.7%	47.8%	89.6%
LH	5.1%	4.1%	9.2%
A	0.9%	0.3%	1.2%
Mar	47.7%	52.3%	

Observed

	M	W	Mar
RH	42.7%	46.8%	89.5%
LH	4.4%	4.8%	9.2%
A	0.6%	0.7%	1.3%
Mar	47.7%	52.3%	

Expected (under independence)

	M	W	Mar
RH	41.7%	47.8%	89.6%
LH	5.1%	4.1%	9.2%
A	0.9%	0.3%	1.2%
Mar	47.7%	52.3%	

	M	W	Mar
RH	42.7%	46.8%	89.5%
LH	4.4%	4.8%	9.2%
A	0.6%	0.7%	1.3%
Mar	47.7%	52.3%	

	M	W
RH	934	1070
LH	113	92
A	20	8

Observed

	M	W
RH	956	1048
LH	98	107
A	13.4	14.6

Expected (under independence)

	M	W
RH	934	1070
LH	113	92
A	20	8

Observed

	M	W
RH	956	1048
LH	98	107
A	13.4	14.6

Expected (under independence)

$$\chi^2 = \frac{(-22)^2}{956} + \frac{22^2}{1048} + \frac{15^2}{98} + \frac{(-15)^2}{107} + \frac{6.6^2}{13.4} + \frac{(-6.6)^2}{14.6} = 11.6$$

2 degrees of freedom

Testing at .01 significance level: Cutoff for χ^2 with 2 degrees of freedom is 9.2. So, we reject the null hypothesis. The difference in handedness rates between the sexes is "highly significant".

$$p\text{-value} = 3 \times 10^{-3}$$

χ^2 test for independence

- * Compute marginal proportions
- * Compute expected proportions by multiplying row proportion by column proportion (a rows, b columns)
- * Compute expected numbers by multiplying by total number
- * Compute chi-squared statistic
- * Compare to chi-squared with $(a-1) \times (b-1)$ degrees of freedom

Example: Diseases linked to blood type?

Example 10.2.2 from Evans and Rosenthal:
X=condition, Y=blood type

condition	O	A	B	Total
peptic ulcer (P)	983	679	134	1796
gastric cancer (G)	383	416	84	883
control (C)	2892	2625	570	6087

Observed

Counts

	O	A	B	Tot
P	983	679	134	1796
G	383	416	84	883
C	2892	2625	570	6087

Observed

Counts

	O	A	B	Tot
P	983	679	134	1796
G	383	416	84	883
C	2892	2625	570	6087

Proportions

	O	A	B	Mar
P	.112	.077	.015	.205
G	.044	.047	.010	.101
C	.330	.299	.065	.694
Mar	.486	.423	.090	

Observed

Expected

Counts

	O	A	B	Tot
P	983	679	134	1796
G	383	416	84	883
C	2892	2625	570	6087

Proportions

	O	A	B	Mar
P	.112	.077	.015	.205
G	.044	.047	.010	.101
C	.330	.299	.065	.694
Mar	.486	.423	.090	

Observed

Expected

Counts

	O	A	B	Tot
P	983	679	134	1796
G	383	416	84	883
C	2892	2625	570	6087

Proportions

	O	A	B	Mar
P	.112	.077	.015	.205
G	.044	.047	.010	.101
C	.330	.299	.065	.694
Mar	.486	.423	.090	

	O	A	B	Mar
P	.100	.087	.018	.205
G	.049	.043	.009	.101
C	.337	.293	.062	.694
Mar	.486	.423	.090	

Observed

Counts

	O	A	B	Tot
P	983	679	134	1796
G	383	416	84	883
C	2892	2625	570	6087

Expected

	O	A	B	Tot
P	874	760	162	1796
G	430	374	79	883
C	2957	2573	547	6087

Proportions

	O	A	B	Mar
P	.112	.077	.015	.205
G	.044	.047	.010	.101
C	.330	.299	.065	.694
Mar	.486	.423	.090	

	O	A	B	Mar
P	.100	.087	.018	.205
G	.049	.043	.009	.101
C	.337	.293	.062	.694
Mar	.486	.423	.090	

Random/Deterministic predictor

- * Our analysis didn't distinguish between predictor and response
- * Problem: The predictor wasn't random. We just picked a certain number
- * Does it matter? (Assume that the individuals picked are still a simple random sample of those with these conditions, or, for controls, of healthy individuals.)

Original computation: n_{ij} = # subjects with condition i , blood type j .

$$\hat{\theta}_{ij} = \frac{n_{ij}}{n} \quad \hat{\theta}_{i.} = \frac{n_{i.}}{n} = \sum_j \hat{\theta}_{ij}$$

$$X^2 = \sum \frac{(n_{ij} - n\hat{\theta}_{i.}\hat{\theta}_{.j})^2}{n\hat{\theta}_{i.}\hat{\theta}_{.j}}$$

degrees of freedom = $ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$

Original computation

$$\hat{\theta}_{ij} = \frac{n_{ij}}{n}$$

$$\hat{\theta}_{i.} = \frac{n_{i.}}{n} = \sum_j \hat{\theta}_{ij}$$

$$X^2 = \sum \frac{(n_{ij} - n\hat{\theta}_{i.}\hat{\theta}_{.j})^2}{n\hat{\theta}_{i.}\hat{\theta}_{.j}}$$

degrees of freedom=
 $ab-1-(a-1)-(b-1) = (a-1)(b-1)$

Deterministic predictor computation

$n_{i.}$ are fixed numbers (of subjects with given conditions)

$$\hat{\theta}_{.j} = \frac{1}{n} \sum_i n_{ij}$$

$$X^2 = \sum \frac{(n_{ij} - n_i\hat{\theta}_{.j})^2}{n_i\hat{\theta}_{.j}}$$

degrees of freedom=
 $a(b-1)-(b-1) = (a-1)(b-1)$

Take-home message: Deterministic categorical predictors are analysed the same way as random predictors.