Chi Squared Tests: Goodness of fit, Independence

9.1, 9.2

Today's topics

Chi Squared Tests

- Goodness of fit (9.1)
- Independence (9.2)
- Exact p-value

χ^2 Goodness of Fit Test - Background

Developed by Pearson in 1900

Tests the appropriateness of different models

Approximate test for use with large samples (large n).

Only appropriate when all expected cell counts are greater than 5. (Otherwise, should consider calculating *exact p-value*

χ^2 Goodness of Fit Test

Random sample of size n is classified into k categories or cells.

- Let $Y_1, Y_2, \ldots, Y_k$ denote the respective cell frequencies; $\sum_{i=1}^{k} Y_i = n$
- Denote cell probabilities $p_1, p_2, \ldots, p_k$.
- H_0: $p_1 = p_{10}$, $p_2 = p_{20}, \ldots,$ $p_k = p_{k0}$.
- H_A: H_0 not true

χ^2 Goodness of Fit Test

	Group 1	Group 2	…	Group k	Total
Observed Freq (O_i)	Y_1	Y_2	…	Y_k	n
Probability | H_o (p_{io})	p_{10}	p_{20}	…	p_{k0}	1
Expected Freq (E_i)	np_{10}	np_{20}	…	np_{k0}	n

χ^2 Goodness of Fit Test

Test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{(Obs_i - Exp_i)^2}{Exp_i} \sim \chi^2_{(k-1)}$$

Reject H_0 if $X^2 \geq \chi^2_{(k-1),\alpha}$

Random Digit Example (Goodness of Fit)

When making random numbers, people are usually reluctant to record the same or consecutive numbers in adjacent positions. Even though these true probabilities should be p_{10} = 1/10 and p_{20} = 2/10, respectively. gt8suv tests a friend's concept of a random sequence by asking them to generate a sequence of 51 *random* digits.

May0 generates the following sequence: (Ex 9.1-1)

5	8	3	1	9	4	6	7	9	2	6	3	0
8	7	5	1	3	6	2	1	9	5	4	8	0
3	7	1	4	6	0	4	3	8	2	7	3	9
8	5	6	1	8	7	0	3	5	2	5	2	

Use a goodness of fit test to make a statistical decision about whether this sequence seems to be truly random or not (at α = 0.05)

5	8	3	1	9	4	6	7	9	2	6	3	0
8	7	5	1	3	6	2	1	9	5	4	8	0
3	7	1	4	6	0	4	3	8	2	7	3	9
8	5	6	1	8	7	0	3	5	2	5	2	

Random Digit Example (Goodness of Fit)

$H_0 : p_1 = p_{10} = 1/10,\ p_2 = p_{20} = 2/10,\ p_= = p_{30} = 7/10$

H_A : ?

	Group 1	Group 2	...	Group k	Total
Observed Freq (O_i)	Y_1 0	Y_2 8	...	Y_3 42	n 50
Probability | H_o (p_{io})	p_{10}	p_{20}	...	p_{k0}	1
Expected Freq (E_i)	np_{10}	np_{20}	...	np_{k0}	n

$$\frac{(0-5)^2}{5} + \frac{(8-10)^2}{10} + \frac{(42-35)^2}{35} = 6.8$$

$$6.8 > 5.991 = \chi^2_{0.05}(2).$$

Definition

Contingency table

Summarizes relationship between categorical variables by displaying frequency distribution.

e.g.

Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
Female	14	4	175	13	4	210
Totals	35	20	320	15	10	400

χ^2 Test for Homogeneity & Independence (9.2)

Tests relationship between two categorical variables.

The difference in the two names depends on how the data is collected.

χ^2 Test for Homogeneity & Independence (9.2)

χ^2 Test for **Homogeneity**: (one margin fixed)

Tests whether two or more sub-groups of a population share the same distribution of a single categorical variable. E.g., do different age groups have the same proportion of people who prefer Twitch, YouTube Live, or Zoom?

Independent random samples from r populations.

Each sample is classified into c response categories.

H_0 : In each category, the probabilities are equal for all r populations.

χ^2 Test for Homogeneity & Independence (9.2)

χ^2 Test for **Independence**: (no margins fixed)

Tests whether two categorical variables are associated with one another in the population, e.g. age group vs video streaming platform preference.

A random sample of size n is simultaneously classified with respect to two characteristics, one has r categories and the other c categories.

H_0 : The two classifications are independent; i.e., each cell probability is the product of the row and column marginal probabilities

(9.2) notes

Homogeneity

Independence

notes

χ^2 Test for Homogeneity & Independence (9.2)

Test statistic (both tests):

$$X^2 = \sum_{cells} \frac{(O-E)^2}{E} \sim \chi^2_{(r-1)*(c-1)}$$

O = observed cell frequency E = $\frac{\text{row total} * \text{column total}}{\text{overall total}}$

(more formally):

$$X^2 = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

χ^2 Test of homogeneity example:

Saumdog is wondering which instructor to select for Stat 420. Albert doesn't know much so he states H_0: their grade distributions are the same.

We collect 2 separate random samples from each instructor and obtain the following data.

	Grade					
Instructor	A	B	C	D	F	Totals
ObeseFuture	8	13	16	10	3	50
Boolean HyperCube	4	9	14	16	7	50

Perform a χ^2 test of homogeneity at significance level 0.05 to determine if these grade distributions are similar.

Instructor	Grade A	B	C	D	F	Totals
ObeseFuture	8	13	16	10	3	50
Boolean HyperCube	4	9	14	16	7	50

$$q = \frac{(8-6)^2}{6} + \frac{(13-11)^2}{11} + \frac{(16-15)^2}{15} + \frac{(10-13)^2}{13} + \frac{(3-5)^2}{5}$$
$$+ \frac{(4-6)^2}{6} + \frac{(9-11)^2}{11} + \frac{(14-15)^2}{15} + \frac{(16-13)^2}{13} + \frac{(7-5)^2}{5}$$
$$= \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} + \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} = 5.18.$$

notes

Example

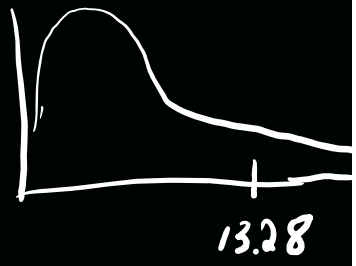
A random sample of 400 undergraduate students at the University of Iowa was selected, then classified by college and gender. Are these variables independent at $\alpha = 0.01$?

Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
Female	14	4	175	13	4	210

Table 9.2-1 Undergraduates at the University of Iowa

Gender	College					Totals
	Business	Engineering	Liberal Arts	Nursing	Pharmacy	
Male	21	16	145	2	6	190
	(16.625)	(9.5)	(152)	(7.125)	(4.75)	
Female	14	4	175	13	4	210
	(18.375)	(10.5)	(168)	(7.875)	(5.25)	
Totals	35	20	320	15	10	400



$$\frac{(21-16.625)^2}{16.625}+\frac{(14-18.375)^2}{18.375}+\cdots+\frac{(4-5.25)^2}{5.25}$$

$$1.15+1.04+4.45+4.02+0.32+0.29+3.69$$

$$+3.34+0.33+0.30 = 18.93.$$

critical value : 13.28

Exact p-value

(not really a new definition)

Review: What is the definition of a p-value?

Calculate this probability directly, instead of using a Normal or Chi Squared approximation.

Exact p-value - Example

Want to test whether a coin is a fair coin or not (loaded in favor of Heads) at α = 0.05.
Data from 100 flips (heads = 1, tails = 0):

```
a = rbinom(100,1
mean(a)  #0.58
```

1 1 1 0 1 0 0 0 1 0 1 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 1 1 0 0 1 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 1 1 0 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 0 1 0 1 0 1

(Could find Z value and do a test that we already know) **OR**

To find exact p-value, just use the definition of the p-value

Let X = # of heads in 100 tosses Under H_0: $X \sim Binomial(100,\ 0.5)$

P-value = $P[X \geq 58]$.

In R:

```
1 - pbinom(57, 100, .6)
```

notes

notes