Ch10:

Pearson χ^2 tests

Ⓐ <u>Goodness-of-fit</u> test

(GoF)

GoF is complicated because it

- leaves much to the imagination (non-P)
- basically a multinomial test (P) under the hood

It can be applied whenever

- your data can be categorized
- & you can calculate expectation for those categories under H_0

Often H_0 involves choosing a distribution (P), so GoF testing is a <u>P/non-P hybrid</u>.

GoF contingency table $(1 \times c)$

↑ # of categories

	class 1	class 2	...	class c	total
obs:	O_1	O_2		O_c	$n = \sum_{j=1}^{c} O_j$

The case of $c=2$ is a Binomial test from Ch 1.

H_0 must boil down to $p_1, \ldots, p_c$ so that we may "model"

$$O \sim \text{Multinom}(n, p)$$

$$\rightarrow e_j = \mathbb{E}\{O_j\} = np_j \quad j = 1, \ldots, c$$

Particular values for p may come from a separate (p) analysis (e.g. MLE), so that the test can be retained as

$$H_0 : p = p_0$$
$$\text{v. } H_1 : p \neq p_0$$

The test stat is

$$e_j = np_j$$

$$\chi^2 = \sum (o_j - e_j)^2 / e_j$$

<u>Math way</u>

It can be shown that

$$\chi^2 \sim \chi^2_{c-1} \quad \text{as } n \to \infty \quad \text{(CLT)}$$

Chapter 10

(B) Homogeneity & Independence tests

r-fold upgrade of multinom/gof test (A)

r populations, c classes/categories

Classification ANOVA (Ch6)

r × c contingency table

1 pop, row class & col class both random

TABLE 10.2: ~~Homogeneity~~ Independence contingency table.

	~~class~~ col 1	~~class~~ col 2	⋯	~~class~~ col c	total
~~pop~~ row 1	o_{11}	o_{12}	⋯	o_{1c}	r_1 $\cancel{n_1} = \sum_j o_{1j}$
~~pop~~ row 2	o_{21}	o_{22}	⋯	o_{2c}	r_2 $\cancel{n_2} = \sum_j o_{2j}$
⋮	⋮	⋮	⋱	⋮	⋮
~~pop~~ row r	o_{r1}	o_{r2}	⋯	o_{rc}	r_r $\cancel{n_r} = \sum_j o_{rj}$
total	$c_1 = \sum_i o_{i1}$	$c_2 = \sum_i o_{i2}$	⋯	$c_c = \sum_i o_{ic}$	$n = \sum_{ij} o_{ij}$

WTK: each pop/row has the same class probabilities or not

$$H_0: O_{ij} \sim p_j \quad \text{for all } i=1,\dots,r,\ j=1,\dots,c$$

$$\text{v. } H_1: \text{some } k \neq j \text{ has } p_{ij} \neq p_{kj}$$

like GoF, each n_j is fixed by design

but each c_j (col sum) is random

note: $n = \sum_{j=1}^{c} c_j = \sum_{i=1}^{r} n_i$

$\downarrow$

ntot

Independence tests are subtly different

WTK: if row/col classifications are <u>independent</u> (H_0) or not (H_1)

$\downarrow$

does knowing a row class determine the col class, or vice versa?

"model" for both (Homogeneity & Indep)

$$O \sim \text{Multinom}(n, P)$$
$\hookrightarrow r \times c$

$$P_{ij} = \frac{n_i}{n} \times \frac{c_i}{n} \quad \text{or} \quad P_{ij} = \frac{r_i}{n} \times \frac{c_i}{n}$$

Homogeneity

Independence

$i = 1, \dots, r \qquad j = 1, \dots, c$

$$\mathbb{E}\{O_{ij}\} = n P_{ij} = \frac{n_i c_i}{n} \equiv e_{ij}$$
$$= \frac{r_i c_i}{n}$$

Test stat is a double-sum upgrade of the GoF one (A)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_{ij} - e_{ij})^2 / e_{ij}$$

Math:

can show $\chi^2 \sim \chi^2_{(r-1)(c-1)}$ as $n \to \infty$