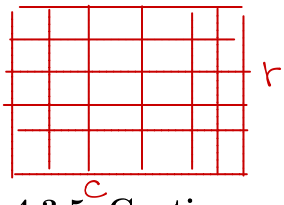

χ^2

$\nu (r-1)(c-1)$

rc X

Lecture 15

4.3.5. Contingency tables. Two major situations.

Suppose we have a data that is presented in the form of a contingency table with r rows and c columns. One may think of at least two illustrative and most commonly met situations where such data may occur:

- To present a tabulation of the data presented in several (r) samples, where the data represent at least a nominal scale with (c) levels of measurement. The purpose is to test the null hypothesis H_0 that the probabilities for each fixed level do **not** differ from sample to sample. Symbolically, if we denote by $p_{ij}, i = 1, 2, \ldots, r; j = 1, 2, \ldots c$ the probabilities for randomly selected value from ith population being classified in the jth class, the null hypothesis says that *all of the probabilities in the same column are equal to each other* ($p_{1j} = p_{2j} = \ldots = p_{rj}$, for all j.)

  Intermezzo: The problem formulated here can be considered a generalisation of the comparison of the proportions of $c = 2$ classes in $r = 2$ populations to the case of r, c being possibly > 2. Let us remaind us that in the case $r = c = 2$, in order to test the hypothesis of equal probabilities, we proceed as follows. Assume that X_1 (out of n_1) and X_2 (out of n_2) are the corresponding observed frequencies in the first category for each of the two samples. We evaluate empirically p_{11} by X_1/n_1 and p_{21} by X_2/n_2. Under $H_0 : p_{11} = p_{21} = p$ we know (thanks to the CLT) that

$$\frac{X_1}{n_1} - p \approx N(0, \frac{p(1-p)}{n_1}), \frac{X_2}{n_2} - p \approx N(0, \frac{p(1-p)}{n_2}) \qquad (1)$$

  Hence

$$X_1 - X_2 \sim \qquad \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1) \qquad (2)$$

  We can now replace in (2) the unknown p via a consistent estimate under H_0, namely the pooled proportion estimate $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ and we get the asymptotic approximation

  pooling!

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1) \qquad (3)$$

$$\boxed{|Z| \overset{?}{>} z_{\alpha/2}}$$

  To test H_0 at level α we can either compare $|Z|$ from (3) with $z_{\alpha/2}$ or, equivalently, compare Z^2 with $\chi^2_{\alpha,1}$. We want to see how the latter procedure is generalised for more than 2 populations and/or more than 2 classes. **End of Intermezzo.**

OR

$$\boxed{Z^2 > \chi^2_{\alpha,1}}$$

$$\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

- For a single sample, each element may be classified into one of r different row-categories according to one criterion and, at the same time, into one of c different column-categories according to a second criterion. The purpose is to test independence of classifications into rows and columns. That is, under the null hypothesis, loosely speaking, the rows and columns represent two independent classification schemes. More precisely, it means that the event "an observation is in row i " is independent of the event "same observation is in column j, for all $i = 1, 2, \ldots, r; j = 1, 2, \ldots c$. Equivalently, we can write that under the null hypothesis, for all $i = 1, 2, \ldots, r; j = 1, 2, \ldots c$, the relation

$$P(\text{ row } i, \text{ column } j) = P(\text{ row } i)P(\text{ column } j)$$

Lung Cancer

		Y	N
S	n_1	O_{11}	O_{12}
N-S	n_2	O_{21}	O_{22}

$N = n_1 + n_2$

must hold.

Both these applications are treated in the same way as far as calculations are concerned. Having in mind the first application, we can think about the element O_{ij} of the table (i.e. its (i,j)th "cell") as the number of observations from the ith sample that fall in jth category. In the second application, each cell frequency is interpreted as the number of observations associated simultaneously with both row i and column j. For the first situation, we can prepare the following table:

	Class 1	Class 2	...	Class c	Totals
Population 1	O_{11}	O_{12}	...	O_{1c}	n_1
Population 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
Population r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Totals	C_1	C_2	...	C_c	N

Obviously, the total number of observations: $N = n_1 + n_2 + \ldots + n_r$. It holds:

$$n_i = O_{i1} + O_{i2} + \ldots O_{ic}, i = 1, 2, \ldots, r$$

for the row totals and

$$C_j = O_{1j} + O_{2j} + \ldots O_{cj}, j = 1, 2, \ldots, c$$

for the column totals.

For the second situation, the following table is appropriate:

	Column 1	Column 2	...	Column c	Totals
Row 1	O_{11}	O_{12}	...	O_{1c}	R_1
Row 2	O_{21}	O_{22}	...	O_{2c}	R_2
...	...	...	...	...	...
Row r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totals	C_1	C_2	...	C_c	N

Note that in the second table, the row totals are denoted by R_j (instead of n_j in the first table to clearly emphasize that the R_j are **random** (whereas n_j are considered as **fixed**).

Coming back to the first situation, assume that the outcomes of the various samples are mutually independent and that each observation can be categorised into exactly one of the c categories. Then under H_0 the expected frequencies within each cell are $E_{ij} = \frac{n_i C_j}{N}$ (i.e. multiply the ith sample size by the proportion $\frac{C_j}{N}$ of all observations in category j.) The idea is to compare the expected and observed frequencies all over the cells by constructing the statistic $Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. For the second situation, denote by $\theta_{ij} = P(\text{row i} \bigcap \text{column j})$ where $\sum_{i=1}^{r}\sum_{j=1}^{c}\theta_{ij} = 1$. Under $H_0 : \theta_{ij} = \theta_{i.}\theta_{.j}$ must hold. The expected frequencies under H_0 are: $E_{ij} = \frac{R_i C_j}{N}$ and the same statistic $Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, with the newly defined E_{ij} can be used. The null distribution of Q in both cases tends asymptotically to $\chi^2_{(r-1)(c-1)}$.

$\theta_{i.} = \sum_{j=1}^{c}\theta_{ij}$

$\theta_{.j} = \sum_{i=1}^{r}\theta_{ij}$

We will give a justification of this fact for the second situation only (the justification for the first situation is similar). Note that if all $\theta_{i.}$ and $\theta_{.j}$ were specified under H_0 then **all** $\theta_{ij} = \theta_{i.}\theta_{.j}$ would be specified. The testing problem would reduce to an ordinary

(# of degrees of freedom without restriction) − (# number of degrees of freedom under the restriction of H_0)

$= (rc-1) \quad -(r+c-2)$

goodness-of-fit test for a composite hypothesis as in 4.3.1. with rc groups and hence the degrees of freedom would have been $(rc-1)$. **Since** we have some freedom to choose $\theta_{i.}$ and $\theta_{.j}$ under H_0, we need to reduce the total $(rc-1)$ df. Careful observation shows that we need to reduce this number by $r+c-2$ (and **not** by $r+c$ (!)) since the two additional restrictions: $\sum_{i=1}^{r}\theta_{i.} = 1 = \sum_{j=1}^{c}\theta_{.j}$ need to also hold true.

The number of free parameters to be estimated under H_0 is therefore $r+c-2$ and we end up with degrees of freedom for the asymptotic distribution equal to $rc-1-r-c+2 = (r-1)(c-1)$.

Incidentally, we notice that the example discussed in the Intermezzo is covered by the general theory when we apply it for $r = c = 2$.

A piece of history. It is the famous Sir Ronald Fisher (1890 – 1962) who first found the correct degrees of freedom for an $r \times c$ contingency table. By contrast, in 1900 Karl Pearson (remember, he was "the father" of chi-square goodness-of fit testing, a very powerful statistician) had argued that for **any application** of his statistic, *df* equalled the number of cells minus 1, or $rc-1$ for two-way tables. When, in 1922, Fisher pointed out in the **Journal of the Royal Statistical Society** that $(r-1)(c-1)$ was the correct value, not surprisingly, Pearson reacted very critically by writing: " I hold that such a view [Fisher's] is entirely erroneous, and that the writer has done no service to the science of Statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society* ... I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us."

Statisticians soon realized that Fisher was correct but he maintained much bitterness over this and other dealings with Pearson. (Other such big dealing was the discussion about the Maximum Likelihood Method's (due to Fisher) superiority in comparison to the method of moments (due to Pearson). We all know now that the MLE is in general superior to the method-of-moments estimator but historically, the fight for advocating the superiority of the MLE has not been easy).

Fisher, of course, did not stay defenceless in these dealings. In 1926, for example, he was able to "dig the knife" into the Pearson family. In his argument, he used 12000 2×2 contingency tables randomly generated (under the hypothesis of independence) by Pearson's own son, E. Pearson (also statistician). Fisher calculated the sample mean of the 12000 χ^2 values for these tables. According to the law of large numbers, it should be very close to the theoretical mean of the chi-square statistic. It turned out to be equal to 1.00001, much closer to the 1.0 predicted by Fisher's formula $(r-1)(c-1) = 1$ than Pearson's $rc - 1 = 3$. In such a way, the dispute was effectively settled. Although the theoretical justification of Fisher's claim came about a decade later (when Wilks derived rigorously the theory of Generalized Likelihoood Testing), Fisher's numerical calculation in 1926 was convincing enough and there was no room for disagreement anymore!

4.3.6. Third situation. Fisher's exact test for a 2×2 contingency tables. The difference between the situations one and two discussed so far is that in the first situation, the marginal row totals are fixed whereas in situation two, none of the marginal row or column totals are fixed. There is also a third situation imaginable in which **both** row and column totals are determined beforehand, and are therefore fixed, not random.

A table that corresponds to this situation, would be:

	Column 1	Column 2	...	Column c	Totals
Row 1	O_{11}	O_{12}	...	O_{1c}	n_1
Row 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
Row r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Totals	c_1	c_2	...	c_c	N

The small c-values c_j indicate that the column totals are also given and not random. The null hypothesis in this case may be either of the hypotheses considered in the two previous situations **under the condition that the row and column totals are fixed.** The test-statistic Q and its limiting distribution remain the same. It should be noted that fixed row and column totals greatly reduce the number of contingency tables possible, and so the **exact** distribution of the Q statistics is more feasible to derive in such situations. When $r = 2$ and $c = 2$, it is indeed possible to derive exact distribution. The test is known as **Fisher's exact test**. It is available in every good contemporary statistical package. Let us discuss shortly this test.

Example (Fisher's (1935) tea taster). A colleague of Fisher claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment that went into Statistics' history as **Fisher's tea taster.** The lady had to taste 8 cups, four had milk added first and the other four had tea added first. She was told there were four cups of each type, so that she would try to select the four that had milk added first. The cups were presented to her in random order. The result of the experiment is shown below:

	Guess Milk poured first	Guess Tea poured first	Total
Milk poured first	3	1	4
Tea poured first	1	3	4
Total	4	4	8

The null hypothesis is that Fisher's colleague's guess was independent of the actual order of pouring; the alternative reflects her claim stating that there is a positive association between true order of pouring and her guess. What could be our conclusion?

Note that both column and row margins are fixed here (not only were the cups served 4 and 4 each **but** the lady also knew in advance that four cups had milk added first). In our terminology n_1, n_2, c_1 and c_2 are fixed and only O_{11} could be considered random (the other values: O_{12}, O_{21}, O_{22} are deterministically determined once O_{11} has been observed). Potential values of O_{11} are (0,1,2,3,4). In general, we would have total of $N = n_1 + n_2$ observations, of which n_1 are of type I, and n_2 are of type II. From these, c_1 have been chosen and we are looking at the probability that O_{11} from them are from type I. This probability is exactly given by the **hypergeometric distribution** and is equal to:

$$\frac{\begin{pmatrix} n_1 \\ O_{11} \end{pmatrix}\begin{pmatrix} n_2 \\ c_1 - O_{11} \end{pmatrix}}{\begin{pmatrix} N \\ c_1 \end{pmatrix}} = \frac{n_1!n_2!c_1!c_2!}{N!O_{11}!O_{12}!O_{21}!O_{22}!}$$

(WHY (!)). In our particular case we get $P(O_{11} = 3) = \frac{4!4!4!4!}{8!1!3!1!3!} = 0.229$. The only table that is more extreme for the alternative, consists of four correct guesses in which case we get $P(4) = \frac{1}{70} = 0.014$. The P-value for the one-sided alternative equals in our case $P(3) + P(4) = .243$ and hence there is not enough evidence against the null hypothesis of independence. We must admit that the test has been very harsh (the sample size is very small), so that the poor lady's only chance against Statistical Significance testing would have been a 100% correct guess that would have left her with a P-values of $0.014 < 0.05$ and we would have accepted her claim of being able to distinguish whether milk or sugar was added first.

Finally, it should be noted that until now we considered *two-way contingency tables.* The observations were classified in two ways, by rows and columns. Of course, one can extend this to include situations where observations are classified according to three or more criteria (i.e. *three- (or more) way contingency tables*). For finer analysis of such multidimensional tables, the theory of **Loglinear Models** has been developed and successfully applied. Within this theory, parameters are introduced to describe the effect of the combination of any levels of the factors in explaining the observed frequencies in the contingency table. Questions about dependence/independence of the classifications can be answered by testing about significance of these parameters but the additional bonus us that when the hypothesis of independence is rejected, one gets a quantitative description of the size of the corresponding effect via the estimated value in the model. Detailed discussion of the theory of Loglinear Models is outside the scope of this course.

4.3.7. Measures of association. Contingency coefficient.

As a measure of the degree of association between families in a contingency table classifying a total of N experimental units, Pearson proposed (as early as in 1904(!)) the **contingency coefficient** C**.** It is defined as $C = \sqrt{\frac{Q}{Q+N}}$. On an intuitive level, it seems quite reasonable to define the coefficient in such a way. It obviously will not exceed the value one (which is good). But a disadvantage of this definition is that the coefficient can never be equal to one.

Many other measures of association will be discussed in Lecture 17.