11 Categorical Data Analysis

In our previous work, we have focused on the analysis of *continuous* and *binary* data covering:

- inferences from a single sample of data:
  One-sample t-test for mean μ, one-sample z-test for proportion p and paired t-test for μ_d.

- inferences from two samples of data:
  Two-sample t-test for difference in means $\mu_1 - \mu_2$ and two-sample z-test for difference in proportions $p_1 - p_2$.

However, there are certain investigations in practice where we collect information as categories and/or counts data. This week, we study a new statistical method and consider experiments where the data are collected on two or more categories.

Motivational Example:

Suppose that the classification of a random sample of 400 workers in a large farm according to their "continent of birth" results in the following count data array corresponding to each of the continents as given below:

Continent of birth	Observed count or frequency
1 Asia	90
2 Europe	75
3 North America	50
4 South America	65
5 Australasia	55
6 Africa	65
Total	400

The workers union may be interested to know whether the proportions of people from each continent are the same. That is to test

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6,$$

where $p_1,\ p_2,\ p_3,\ p_4,\ p_5$ and p_6 are the true proportions of workers from six continents.

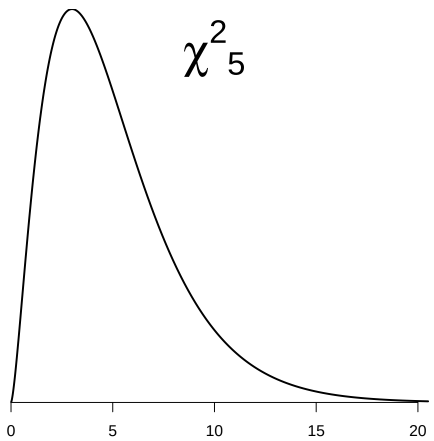
Note:

In the above case, the union is interested in testing the hypothesis on categorical data/variables. It is clear that this is a generalization of binary variables with more classes. Therefore, this topic, known as *categorical data analysis*, is very popular in many scientific research areas.

11.1 Analysis of Categorical Data

The analysis of such categorical data is based on the properties of another continuous distribution called the *Chi-square* distribution, denoted by χ^2. This distribution is also indexed by a single parameter ν or k for the *degrees of freedom* (df).

A typical shape of a Chi-square distribution is given below:



Properties of the Chi-Square Distribution

1. This is a *continuous* distribution taking only *positive* values.

2. This is a *right-skewed* distribution in general. The distribution becomes less skewed as the df increases.

3. This is the distribution for the sum of a number (say ν) of *independent squared standard normal* random variables. The number ν gives the df of the distribution.

4. The Chi-square table gives the percentage points of Chi-square distributions for various df and *right tail area* (or the probability), similar to the t-table for t-distribution.

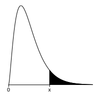


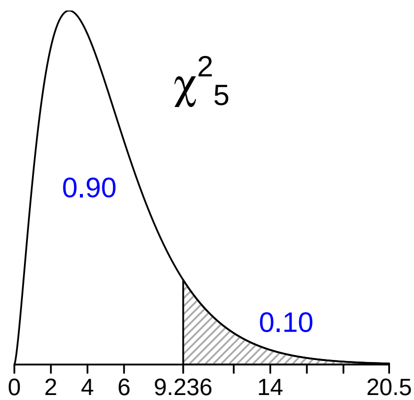
Table 3: Chi-square Distribution Table

Percentage point $P(\chi^2_\nu > x) = p$ for the χ^2 distribution with ν degrees of freedom.

p	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
ν								
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
90	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

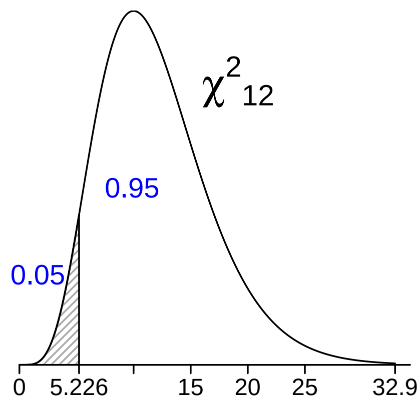
Example: 1. Shade the region for $P(\chi_5^2 \geq 9.236)$ and find this probability.

Solution: Across the row with df=5 in the Chi-square table:
$P(\chi_5^2 \geq 9.236) =$ ____ or $P(\chi_5^2 \leq 9.236) =$ ____.



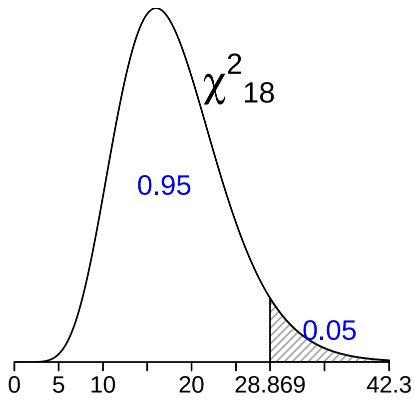
Example: 2. Shade the region $P(\chi_{12}^2 \leq 5.226)$ and find the corresponding probability.

Solution: Across the row with df=12 in the Chi-square table:
$P(\chi_{12}^2 \leq 5.226) =$ ______________.



Examples 3. Find $P(\chi^2_{18} > 28.869)$.

Solution: Across the row with df=18 in the Chi-square table: $P(\chi^2_{18} > 28.869) =$ ____.

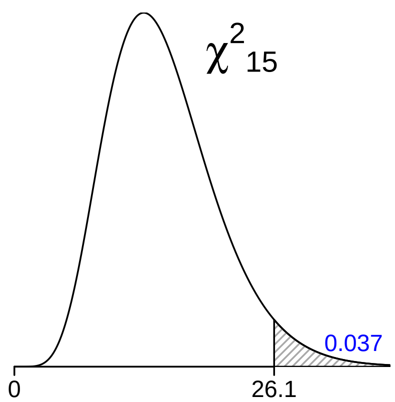


Note: Since the chi-square distribution is a continuous distribution,

$$P(\chi^2_{18} > 28.869) = P(\chi^2_{18} \geq 28.869) = 0.05.$$

Example: 4. Find the lower and upper bound for $P(\chi^2_{15} > 26.1)$.

Solution: Now it is clear that $P(\chi^2_{15} > 26.1)$ is in the interval ____________ and therefore $P(\chi^2_{15} > 26.1)$ is a small probability.



Note that the exact probability 0.037 can be obtained using the R command `1-pchisq(26.1,15)`.

11.2 Chi-square Tests

In this course, the Chi-square test is applied to determine:

1. How well the given set of categorical data fit to a theoretical (or a hypothetical) model. This is known as the *Chi-square goodness-of-fit* (GOF) test.

2. Whether there exists an association between two categorical variables (in contingency tables). This is related to the analysis of Contingency Tables.

11.2.1 Chi-square Goodness-of-Fit Test (P.178-181; omit P.156-173)

Example: Suppose that a psychologist is interested in determining whether mentally retarded children, given a choice of four colours, prefer one colour over the other. The researcher conjectures that colour preference may have some effect on behaviour. Eighty mentally retarded children are given a choice of brown, orange, yellow, or green T-shirts. This is a tally of their selection:

Colour	Frequency
Brown	25
Orange	18
Yellow	19
Green	18
Total	80

Do the children have a colour preference?

Solution: The numbers appeared on this table are called observed frequencies and are denoted by O_i.

In our case: $O_1 = 25, \quad O_2 = 18, \quad O_3 = 19, \quad O_4 = 18,$

1. Firstly, we set up the following hypotheses:

H_0: there is no colour preference, i.e. ____________________ vs
H_1: there is a colour preference, i.e. ____________________

Under the null hypothesis, how many values do we expect in each category?

One would expect $\frac{1}{4}$ of 80, i.e. $E_i = np_{i0} =$ ____________ of children to select each colour under H_0 of no colour preferences. These expected frequencies are denoted by E_i.

$E_1 = 20, \quad E_2 = 20, \quad E_3 = 20, \quad E_4 = 20$

2. **Test statistic:** If the null hypothesis is true, we expect the observed and expected frequencies to be close to each other. In other words, their differences should be small. In this example, they are:

$$O_1 - E_1 = 25 - 20 = 5, \quad O_2 - E_2 = 18 - 20 = -2,$$
$$O_3 - E_3 = 19 - 20 = -1, \; O_4 - E_4 = 18 - 20 = -2$$

However they are canceled when summed over categories. To avoid cancellation, the differences are squared:

$$(O_1 - E_1)^2 = 5^2 = 25, \quad (O_2 - E_2)^2 = (-2)^2 = 4,$$
$$(O_3 - E_3)^2 = (-1)^2 = 1, \; (O_4 - E_4)^2 = (-2)^2 = 4$$

To facilitate comparison, these squared differences need to be standardized to eliminate the scale effect. An obvious way is to divide the squared differences by their expected values:

$$\frac{(O_1-E_1)^2}{E_1}=\frac{25}{20}=1.25,\quad \frac{(O_2-E_2)^2}{E_2}=\frac{4}{20}=0.20,$$
$$\frac{(O_3-E_3)^2}{E_3}=\frac{1}{20}=0.05,\quad \frac{(O_4-E_4)^2}{E_4}=\frac{4}{20}=0.20.$$

Then the sum is 1.70 and it gives a *measure of overall fit* between the observed and expected counts across categories under the null hypothesis. Hence the sum serves as the test statistic for the χ^2 GOF test and is given by:

$$X^2_{\text{obs}}=\sum_{i=1}^{g}\frac{(O_i-E_i)^2}{E_i}\sim\chi^2_{g-1}.$$

It is clear that the large value of X^2_{obs} or simply X^2_0 will argue against H_0, in favour of H_1. We need a distribution to check if X^2_0 is large to indicate inconsistency of data with H_0.

Since X^2_0 is the sum of a number of squares for the standardized residuals or differences, $d_i=\frac{O_i-E_i}{\sqrt{E_i}}$, it follows a χ^2 distribution with df$=g-1$ where g denotes the number of classes.

The above calculation can be performed using the following table:

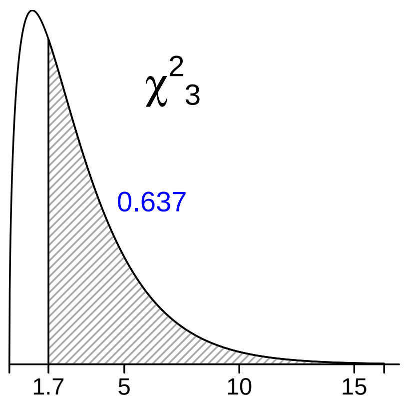
Colour	Observed O_i	Expected E_i	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
Brown	25	__	_	__
Orange	18	__	__	_____
Yellow	19	__	__	_____
Green	18	__	__	_____
Total	80	__	_	__

Then the test statistic is:

$$X_0^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \underline{\hspace{3cm}}.$$

3. *P*-**value:** Since $g = 4$, we have df $= 3$. Therefore, the corresponding P-value is given by:

$$P\text{-value} = \underline{\hspace{5cm}}.$$



4. **Conclusion:** ____

________. The mentally retarded children have no significant preference with respect to the four colours.

In general, with the observed frequencies $x_1, x_2, ..., x_g$ from g groups, a model (a probability distribution):

$$p_1 = p_{10}, \; p_2 = p_{20}, \; \cdots \; , \; p_g = p_{g0},$$

where $p_{i0} > 0$ and $\sum_{i=1}^{g} p_{i0} = 1$, provides a good fit to the observations x_i if the test statistic

$$X_0^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{g} \frac{(x_i - np_{i0})^2}{np_{i0}} \sim \chi^2_{g-1}$$

is small where $n = \sum_{i=1}^{g} x_i$ is the sample size.

The P-value is $P(\chi^2_{g-1} \geq X_0^2)$.

Notes:

1. We don't do "two times the probability" for this P-value because the test statistics X^2_{obs} is always *one-sided* as large positive and negative r_i will both give large X^2_{obs}.

2. The formula for X^2_{obs} is given in the formulae sheet. If there are g groups in the problem, then the df is $g-1$ (one less than the total number of groups).

3. The assumptions are that each expected frequency is $E_i = np_{0i} \geq 5$. If there are categories with $E_i < 5$, then adjacent categories should be combined and the new df$=g'-1$ where g' is the new number of categories.

Example: In an experiment involving a dihybrid cross of flies, 144 progeny were classified by phenotype as follows.

AB	Ab	aB	ab	Total
86	30	23	5	144

Genetic theory predicts a ratio 9:3:3:1 for AB:Ab:aB:ab. Do the data support the theory?

Solution: The χ^2 GOF test for proportions is

1. Hypotheses: ______________________________ vs

   ________________________.

2. Test statistic: The calculation of the expected frequencies under the null hypothesis H_0, say,

   $E_1 = np_{10} =$ ____________ from the group AB,
   $E_2 = np_{20} =$ ____________ from the group Ab and so on

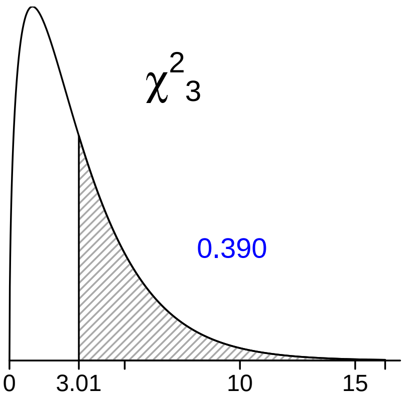
   are performed by completing the following table:

Type	Obs. O_i	Exp. $E_i = np_{i0}$	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
AB	86	________	________	________
Ab	30	________	________	________
aB	23	________	________	________
ab	5	________	________	________
Total	144	___	___	$X_0^2 =$ ____

Hence the test statistic is:

$$X_0^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} = \_\_\_\_$$

3. P-value: __________________.



4. Conclusion: ____ ________. We conclude that the data fit well the given model.

Example: (2008 June Exam) Mendellian inheritance predicts that the ratio of red, white and pink should be 1:1:2 in cross-pollination. A biologist wanted to test this claim and counted the number of red, white and pink flowered plants resulting after cross pollination of 260 white and red sweet peas. The results were:

Colour	Red	White	Pink	Total
Number	72	63	125	260

Test the null hypothesis that the model fits well for the data.

Solution:

1. Hypotheses: ______________________ vs ______________________.

2. Test statistic: Under H_0, we have

Colour	Observed O_i	Expected E_i	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
Red	72	______	______	______
White	63	______	______	______
Pink	125	______	______	______
Total	260	___	___	___

$X_0^2 =$ ______________________________________

3. P value: ______________________________

4. Conclusion: ______________________________ with H_0. The ratio of red, white and pink flowered plants is 1:1:2 in cross-pollination.

11.2.2 Chi-square test for testing independence of two categories (P.173-177)

Chi-square test can be applied to *contingency tables* for testing independence of two categories.

Definition: A *contingency table* containing r rows (categories) and c columns (categories) of frequencies on two different categorical variables is called an $r \times c$ contingency table or a two-way table. It displays information on two categorical variables.

An Illustrative Example
A random sample of 100 women who have had a child within the past year are classified by whether or not they receive nutritional counselling and whether or not they are breastfeeding their child. The results are:

	Nutritional counselling		
Breastfeeding	Yes	No	Row total r_i
Yes	30	21	51
No	18	31	49
Col. total c_j	48	52	100

Note: In this data matrix, each box is called a *cell* and there are 4 cells altogether, from 2 rows and 2 columns. This table is known as a 2×2 contingency table.

Let O_{ij} be the observed frequency in the box in row i and column j and r_i and c_j denote the i-th row total and j-th column total respectively. Therefore the data matrix is:

$$\begin{array}{|c|c|}\hline O_{11} & O_{12} \\ \hline O_{21} & O_{22} \\ \hline \end{array} = \begin{array}{|c|c|}\hline 30 & 21 \\ \hline 18 & 31 \\ \hline \end{array}$$

The χ^2 test for independence between two categories is:

1. Hypotheses: ______________________________ vs
_________________________________.

   Let p_{ij} be the probability that an observation comes from cell (i, j). Recall that if events A and B are independent,

   $$P(A \cap B) = P(A)P(B).$$

   Hence the hypotheses can be rewritten as:

   ______________________________ vs

   ____________________.

2. Test statistic:
   To derive the test statistic, we first calculate the expected frequency $E_{ij} = np_{ij} = np_i \times p_j$ in each cell assuming "Nutritional Counselling" and "Breastfeeding" are independent under H_0. These E_{ij} are estimated by:

   $$E_{ij} = n\hat{p}_{ij} = n\hat{p}_i \times \hat{p}_j = n\,\frac{r_i}{n} \times \frac{c_j}{n} = \frac{r_i \times c_j}{n}$$

   The calculation of E_{ij} for the data is illustrated below:

$$\begin{array}{|c|c|}\hline E_{11} & E_{12} \\ \hline E_{21} & E_{22} \\ \hline\end{array} = \begin{array}{|c|c|}\hline \frac{r_1 \times c_1}{n} & \frac{r_1 \times c_2}{n} \\ \hline \frac{r_2 \times c_1}{n} & \frac{r_2 \times c_2}{n} \\ \hline\end{array} = \begin{array}{|c|c|}\hline \quad & \quad \\ \hline \quad & \quad \\ \hline\end{array} = \begin{array}{|c|c|}\hline \quad & \quad \\ \hline \quad & \quad \\ \hline\end{array}$$

   If the variables are independent, then the observed and expected frequencies must be close to each other. Therefore the test statistic is the sum of all squared residuals as before:

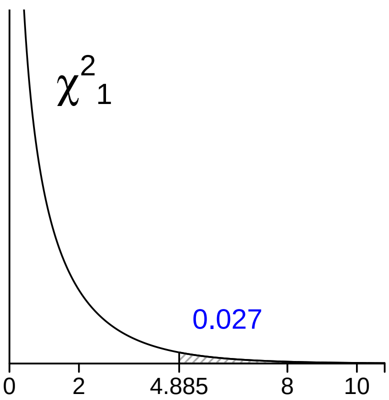
$$X^2_{\text{obs}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-\frac{r_i\times c_j}{n})^2}{\frac{r_i\times c_j}{n}} \sim \chi^2_{(r-1)(c-1)}$$

where x_{ij} is the observed value of O_{ij} and the distribution of χ_{obs} is approximately χ^2 with $(r-1)(c-1)$ df.

Hence we calculate X^2_{obs} for the above contingency table:

$$X_0^2 = \underline{\hspace{10cm}}$$
$$= \underline{\hspace{1cm}}$$

3. P-value: ______________________________



4. Conclusion: ________________________________ ________________________. The two variables, Nutritional Counselling and Breastfeeding, are dependent.

Example: Each member of a sample of 166 persons taking a medical test on blood glucose level (BGL) was classified by

(i) whether or not he/she pass the test on BGL and
(ii) socioeconomic level (the higher the score, the higher the level) as follows:

	Socioeconomic level				
BGL results	1	2	3	4	5
Passed	2	13	35	40	40
Failed	1	7	15	6	7

Using a suitable Chi-square test determine whether the two variables are independent.

Solution:

1. Hypotheses:

   ____
   __________________________.

2. Test statistic:

   In the given 2×5 contingency table, the frequencies at level 1 of socioeconomic status are smaller than 5. Therefore the Chi-square test may not give a satisfactory result. To resolve this, the theory suggested to combine the levels 1 and 2 to obtain a reduced 2×4 contingency table as below:

	Socioeconomic level				
BGL results	1 or 2	3	4	5	Total
Passed	15	35	40	40	130
Failed	8	15	6	7	36
Total	23	50	46	47	166

Then we calculate the expected frequencies under H_0 as:

$E_{11}=$	$E_{12}=$	$E_{13}=$	$E_{14}=$
$E_{21}=$	$E_{22}=$	$E_{23}=$	$E_{24}=$

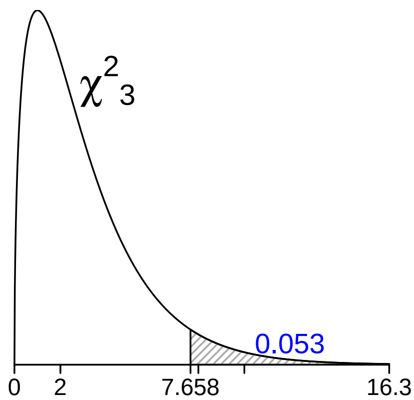
and the squared standardized differences d_{ij}^2 are:

$d_{11}^2=$	$d_{12}^2=$	$d_{13}^2=$	$d_{14}^2=$
$d_{21}^2=$	$d_{22}^2=$	$d_{23}^2=$	$d_{24}^2=$

Hence the test statistic is

$$X_0^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \underline{\hspace{8cm}}.$$

3. P-value: ______________________________



4. Conclusion: ____________________________
________. That is, these two variables can be considered as independent.

Read example 10.33 on P.173, example 10.34 on P.174-176 and example 10.35 on P.177.