Scope

This document shows the solution for the data set in the table below. Your data set will almost certainly be different than this one, but the methods illustrated here are still valid for your data set. To refresh your memory, we begin with the general formulas involved. If you want to skip the formulas, click here.

Formulas

There are n observations, each of which can be classified into one of c categories. Y_i is the number of observations classified into category i. The values Y_i are displayed in a one-way table. It is assumed that these observations are a random sample from a population having a proportion p_i in category i.

Hypotheses

It is desired to test the hypotheses

$$\begin{aligned} H_0: &\quad p_i = p_i^0,\ i = 1, \ldots, c \\ H_a: &\quad p_i \neq p_i^0,\ \text{for at least one } i,\ i = 1, \ldots, c, \end{aligned}$$

for pre-specified values $p_1^0, \ldots, p_c^0$.

Expected Values

If H_0 is true, the expected number of observations in category i is $np_i^0,\ i = 1, \ldots, c$.

Test Statistic

The test statistic is

$$X^2 = \sum_{i=1}^{c} \frac{(Y_i - np_i^0)^2}{np_i^0}$$

Hypothesis Test

If H_0 is true, X^2 has approximately a χ^2 distribution with $c-1$ degrees of freedom. The p-value is given by $P(W \geq x^{2*})$, where $W \sim \chi^2_{c-1}$ and x^{2*} is the observed value of X^2.

Data

There are $n = 2000$ lottery winners in the data set distributed among $c = 5$ age groups as shown by the Y_i values in the table below. The hypothesized population proportions in each age group p_i^0 are in the next column followed by the expected numbers $(2000p_i^0)$ in those groups.

Age group (i)	Number (Y_i)	p_i^0	np_i^0	$\frac{(Y_i-np_i^0)^2}{np_i^0}$
Younger than 20	262	0.15	300	$4.81\overline{33}$
20 to 30	301	0.15	300	$0.00\overline{33}$
30 to 40	326	0.15	300	$2.25\overline{33}$
40 to 50	311	0.15	300	$0.40\overline{33}$
Older than 50	800	0.40	800	0.0000
				$X^2 = 7.47\overline{33}$

SAS Code

The following SAS code will produce all the output needed to answer questions a.)-e.) (and more). The data step reads the data into the SAS data set *salary*. The SAS procedure *proc reg* performs the regression and outputs results. The corr option produces a correlation matrix.

```
data lottery;
 input group $ 1-16 number;
 datalines;
Younger than 20 262
20 to 30        301
30 to 40        326
40 to 50        311
Older than 50   800
;
run;

proc freq data=lottery order=data;
 tables group/nocum chisq testp=(15 15 15 15 40);
 weight number;
 ods output OneWayFreqs=table_out;
run;

data table_out;
 set table_out;
 Expected=2000*TestPercent/100;
 drop Table F_group;
run;
title 'Frequency Table with Expected Values';
proc print data=table_out noobs;
run;
```

Solutions

Below, we show how the SAS output can be used to answer questions a.)-e.).

a.) This is the last column in the following output table:

Frequency Table with Expected Values

group	Frequency	Percent	Test Percent	Expected
Younger than 20	262	13.10	15.00	300
20 to 30	301	15.05	15.00	300
30 to 40	326	16.30	15.00	300
40 to 50	311	15.55	15.00	300
Older than 50	800	40.00	40.00	800

b.) Clear.

c.)-d.) The test statistic value, $X^2 = 7.4733$ and p-value, 0.1129, are found in the following output table:

Chi-Square Test
for Specified Proportions

Chi-Square	7.4733
DF	4
Pr > ChiSq	0.1129

e.) Since the p-value is greater than $\alpha = 0.01$, we do not reject the researchers claim.