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 $r \times c$ categories. A summary table consisting of r rows and c columns (called a *two-way table*) is used to display the numbers of observations in each category. Y_{ij} is the number of observations classified into the row i and column j category, which we'll call category ij. In the present example, it is assumed that these observations are a random sample from a population having a proportion p_{ij} in category ij.

We also define the marginal totals $Y_{i\cdot} = \sum_{j=1}^{c} Y_{ij}$ and $Y_{\cdot j} = \sum_{i=1}^{r} Y_{ij}$, which are the total number of observations in row i and column j, respectively.

Hypotheses

It is desired to test the hypotheses that the two ways of classifying the observations are independent. The following null hypothesis expresses the independence assumption. the alternative hypothesis is that the ways of classifying are not independent.

$$\begin{aligned} H_0: &\quad p_{ij} = p_{i\cdot}p_{\cdot j},\ i = 1, \ldots, r,\ j = 1, \ldots, c \\ H_a: &\quad p_{ij} \neq p_{i\cdot}p_{\cdot j},\ \text{for at least one pair } ij,\ i = 1, \ldots, r,\ j = 1, \ldots, c, \end{aligned}$$

where $p_{i\cdot}$ and $p_{\cdot j}$ are the marginal probabilities

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij},\ p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

Expected Values

If H_0 is true, the expected number of observations in category ij is $np_{i\cdot}p_{\cdot j},\ i = 1, \ldots, r,\ j = 1, \ldots, c.$ Since we do not know the true values $p_{i\cdot}$ and $p_{\cdot j}$ we estimate them as

$$\hat{p}_{i\cdot} = \frac{Y_{i\cdot}}{n},\ \text{and } \hat{p}_{\cdot j} = \frac{Y_{\cdot j}}{n}.$$

We then estimate the expected value in category ij as $n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$.

Test Statistic

The test statistic is

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}.$$

Hypothesis Test

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

Data

In order to study the relationship between living environment and political party registration, a researcher took random samples of urban, suburban and rural registered voters. She hypothesized that living environment and political party registration are independent. The data are in the table below.

	Registration			
Environment	Democrat	Republican	Independent	Total
Urban	38	12	18	68
Suburban	24	26	32	82
Rural	4	17	29	50
Total	66	55	79	200

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 env_pol;
 input environment $ @;
 do i=1 to 3;
  input registration $ number @@;
  output;
 end;
 drop i;
datalines;
urban democrat 38 republican 12 independent 18
suburban democrat 24 republican 26 independent 32
rural democrat 4 republican 17 independent 29
;
run;
proc freq data=env_pol order=data;
 tables environment*registration/  chisq expected nopercent norow nocol;
 weight number;
run;
```

Solutions

a.) This can be answered from the SAS output below. The second entry in each cell is the expected value.

Table of environment by registration

environment registration

Frequency Expected	democrat	republic	independ	Total
urban	38	12	18	68
	22.44	18.7	26.86	
suburban	24	26	32	82
	27.06	22.55	32.39	
rural	4	17	29	50
	16.5	13.75	19.75	
Total	66	55	79	200

If computing by hand, the expected value for category urban/republican, for example, is $68 \times 55/200 = 18.7$.

b.) and c.) These answers can also be found in the SAS output. SAS gives a whole lot of tests, but you need only use the chi-square test. On the output it looks like this:

Statistic	DF	Value	Prob
Chi-Square	4	31.5612	<.0001

The answer for b.) is the value of X^2, test statistic, $x^{2*} = 31.5612$. Its p-value is less than 0.0001 (SAS doesn't like to report smaller values). Not to worry, though: WeBWorK will accept 0.0001 as a correct answer for c.).

d.) This problem sets the level $\alpha = 0.1$ for significance of the test (your problem may set a different level). Since the p-value is less than than $\alpha = 0.1$, we reject the null hypothesis and conclude environment and registration are not independent.