**Lab 2**

1. HW1.txt
   1. Summary

The summary of a data set can be viewed in R by storing a data table to a variable and then calling that variable:

> summary(HW1)

Gender Smoking CreditCard

Female:248 Never smoked:131 American Express: 54

Male :252 Quit :243 Master Card :209

Smokes :126 Other : 5

Visa :232

Age Belief\_in\_afterlife

<25 : 35 N:252

>45 :118 Y:248

25 to 35:154

35 to 45:193

* 1. 95% CI for Visa

Confidence intervals for a proportion can be found in R using the proportion test (prop.test) command (the binom.test could also be used). This can be done providing the function with the number of successes (in this case, the number of Visa holders, found in column 4 of the table of card holder counts), and the total number of cards/cardholders considered, (the sum of all numbers of cards held). While the function is better used for hypothesis tests, implicit in it is a return of a 95% confidence interval for the proportion in question. The formula used to compute the confidence interval is

The following result is found:

> prop.test(card[4], sum(card))

1-sample proportions test with continuity correction

data: card[4] out of sum(card), null probability 0.5

X-squared = 2.45, df = 1, p-value = 0.1175

alternative hypothesis: true p is not equal to 0.5

**95 percent confidence interval:**

**0.4197422 0.5088193**

sample estimates:

p

0.464

* 1. Test the hypothesis that the true proportion of males is 50%

This time, the binom.test function was used. This function conducts a hypothesis test on a binomial variable in the same manner prop.test does. A hypothesis test is conducted given the number of successes, the number of trials, the hypothesized proportion (in this case 0.5), and which area of the distribution to reject (in this case both tails comprise the rejection region). A test statistic is computed using . The probability of a result as extreme or more extreme resulting for is then computed using the null distribution based on p = 0.5. This is the P-value, which is used to guide our decision:

> binom.test(gender[2], sum(gender), p=.5, alternative='t')

Exact binomial test

data: gender[2] and sum(gender)

number of successes = 252, number of trials = 500, **p-value =**

**0.8933**

alternative hypothesis: true probability of success is not equal to 0.5

With a p-value greater than 0.10, I cannot reject the null hypothesis that n is not = 0.5 at the 90% confidence level. I conclude that the true proportion of males likely is 50%.

* 1. Use the chi-square test to check whether the true smoking proportions are [Smokes, Quit, NeverSmoked] = [25%, 50%, 25%]

The chi-squared test is traditionally used as a test for independence, however, it can be used to test the likelihood of other expected proportional distributions in contingency tables. The traditional chi-squared statistic is computed using , where , and then the result is used to find the total probability of the observed result and any more extreme than the observed result in using the chi-squared distribution, given the degrees of freedom, which is found by

. In this case however, R is allowing us to input our own expected values, [.25, .5, .25], rather than using the traditional expected values based on independence of observations:

> chisqCTsmoke<-chisq.test(smoke,p=c(.25,.5,.25))

> chisqCTsmoke

Chi-squared test for given probabilities

data: smoke

X-squared = 0.492, df = 2, **p-value = 0.7819**

The P-value of 0.7819 suggests that the null hypothesis should not be rejected, the null hypothesis being that the true proportions of [Smokes, Quit, NeverSmoked] are [25%, 50%, 25%]. This conclusion would hold true for any reasonable confidence level.

* 1. Construct CTs of Gender vs. Smoking, Gender vs. Belief in Afterlife, and Age vs. Belief in Afterlife

Contingency Tables are simply constructed using the xtabs command and stored to a variable in R:

* + 1. Gender vs. Smoking

> GvS

Smoking

Gender Never smoked Quit Smokes

Female 33 125 90

Male 98 118 36

* + 1. Gender vs. Belief in Afterlife

> GvB

Belief\_in\_afterlife

Gender N Y

Female 98 150

Male 154 98

* + 1. Age vs. Belief in Afterlife

> AvB

Belief\_in\_afterlife

Age N Y

<25 16 19

>45 71 47

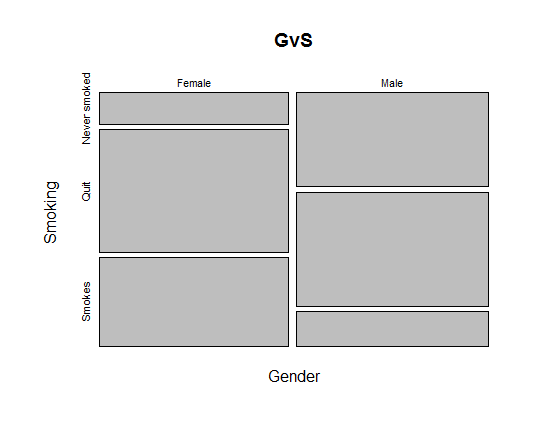
25 to 35 74 80

35 to 45 91 102

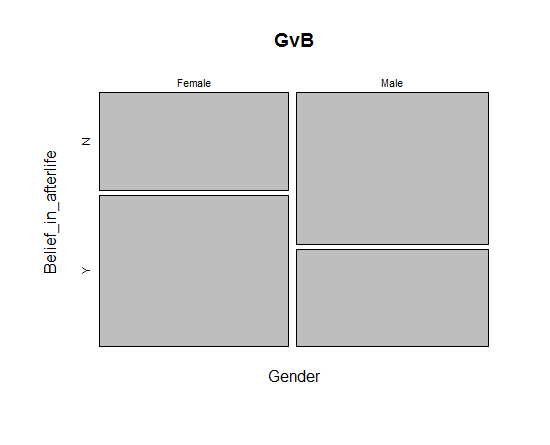
* 1. Use mosaicplot() to plot the CTs from item (e)

Mosaic plots illustrate the proportions of observations found in cross-classifications of variables:

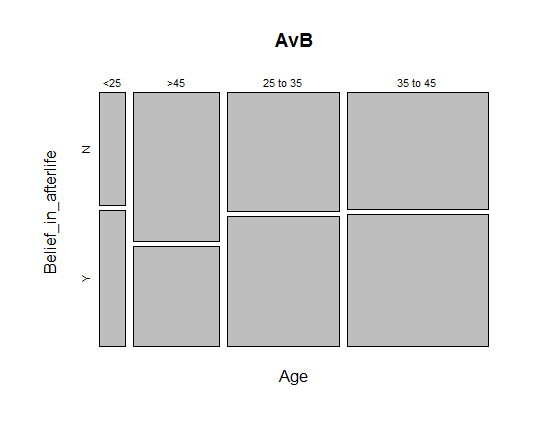
* + 1. Gender vs. Smoking



* + 1. Gender vs. Belief in Afterlife



* + 1. Age vs. Belief in Afterlife



* 1. Test for the existence of associations in the above CTs using different tests (chi-square, proportion, Fisher); compare p-values, use residuals to make conclusions.
     1. Gender vs. Smoking

The chi squared test was used once more; the theory was described above in #1 part d.

Pearson's Chi-squared test

data: GvS

X-squared = 55.568, df = 2, p-value = 8.582e-13

> csGvS$residuals

Smoking

Gender Never smoked Quit Smokes

Female -3.9668671 0.4073407 3.4791227

Male 3.9352581 -0.4040949 -3.4514002

> csGvS$stdres

Smoking

Gender Never smoked Quit Smokes

Female -6.5043545 0.8003151 5.6663537

Male 6.5043545 -0.8003151 -5.6663537

The infinitesimal P-value of 0.0000000000008582 makes it easy to reject the null hypothesis that Gender and Smoking are independent from one another. The residuals, both normal and standardized, are large enough to suggest that Gender and Smoking status are not independent of eachother.

* + 1. Gender vs. Belief in Afterlife

The Fisher Exact Test directly computes the probability of observing a certain distribution of observations within a contingency table, or any result more extreme. This is done for each probability found using to find the probability of observing each result and those more extreme, giving us a P-value with which to test the null hypothesis that the variables in question are independent:

Fisher's Exact Test for Count Data

data: GvB

**p-value = 9.848e-07**

alternative hypothesis: true odds ratio is less than 1

95 percent confidence interval:

0.0000000 0.5713993

sample estimates:

odds ratio

0.4165117

Given such a small P-value, 0.0000009848, it is easy to reject the null hypothesis that Gender and Belief in Afterlife are independent. In fact, it would appear that the odds of a female believing in the afterlife are roughly two times that of a male believing in the afterlife.

For practice, the proportion test was also used. The prop.test function uses a chi-squared test to compare values in a two column matrix, based on the assumption that the proportions of the entries in the different columns will be the same

> propGvB

2-sample test for equality of proportions with

continuity correction

data: GvB

X-squared = 22.4613, df = 1, **p-value = 2.144e-06**

alternative hypothesis: two.sided

95 percent confidence interval:

-0.3055361 -0.1263636

sample estimates:

prop 1 prop 2

0.3951613 0.6111111

The resulting P-value of 0.000002144 would suggest that we should reject the null hypothesis, that the proportions of males believing in the afterlife are similar to that of females, with a high level of confidence.

1. Age vs. Belief in Afterlife

Again, the chi-squared test is used to test for independence:

Pearson's Chi-squared test with simulated p-value (based on

5000 replicates)

data: AvB

X-squared = 5.9676, df = NA, p-value = 0.1188

> chisSimAvB$stdres

Belief\_in\_afterlife

Age N Y

<25 -0.5749264 0.5749264

>45 2.4283438 -2.4283438

25 to 35 -0.7005819 0.7005819

35 to 45 -1.1523572 1.1523572

The P-value is not quite small enough, 0.1188, to reject the null hypothesis that belief in afterlife and age are independent at the 90% confidence level. The majority of the standard residuals have an absolute value of less than 2, suggesting that these data are independent. I conclude that the data are not associated.

1. HW2.txt. The data is hypothesized to be produced by a phenomenon that follows a Binomial distribution with parameters (n = 6, p = 0.7).
   1. Use the Chi-square test to decide whether data confirms the hypothesis about the Binomial distribution with the following parameters ( 6, 0.7 )

The following tests were performed by using the chisq.test function to evaluate how well the data matched the ideal outcome of X ~ Bin( 6, 0.7), by passing the chisq.test function parameters that would replace the traditional expected counts in a 1 by 6 matrix with one that would display the counts of successes in the 6 trial, 10,000 case sample to be expected if the counts were based on a binomial sample with probability of success p=0.7. The P-value of this test would then ideally indicate whether or not it is acceptable to assume that the population the data sets come from

* + 1. V1

> csV1<-chisq.test(V1, p=dbinom(seq(0,length(V1)-1), length(V1)-1, 0.7))

> csV1

Chi-squared test for given probabilities

data: V1

X-squared = 5.0984, df = 6, **p-value = 0.5312**

With a P-value of 0.5312, it is very unlikely that the distribution of the data in V1 are distributed in the same manner as a Binomial distribution with p = 0.7. I conclude that this data does not have such a distribution, not supporting the hypothesis.

* + 1. V2

> csV2<-chisq.test(V2, p=dbinom(seq(0,length(V2)-1), length(V2)-1, 0.7))

> csV2

Chi-squared test for given probabilities

data: V2

X-squared = 16.0166, df = 6, **p-value = 0.01367**

In the V2 data set, our test for association produces a P-value of 0.01367, indicating with about 98.5% confidence that the distribution of the population V2 came from is the same as that of a Binomial distribution with probability of success p = 0.7. I conclude that V2 was sampled from such a distribution, supporting the hypothesis that the data was produced by the mentioned distribution..

* + 1. V3

> csV3<-chisq.test(V3, p=dbinom(seq(0,length(V3)-1), length(V3)-1, 0.7))

> csV3

Chi-squared test for given probabilities

data: V3

X-squared = 0.4526, df = 6, **p-value = 0.9984**

The test of V3 produces a very high P-value, such that it is difficult to believe that the data was randomly drawn from a sample that resembled the aforementioned binomial distribution. This P-value warrants that the null hypothesis that the data of V3 ~ Bin (6, 0.7) fail to be rejected.

* 1. For 653 students.