MONT 107N – Understanding Randomness
Selected Solutions for Problem Set 5
April 27, 2010

Chapter 28/3. Reading between the lines a bit, it can be seen that the question being asked here is equivalent to asking whether marital status and labor force status are *independent*. Hence we can answer the question by doing a χ^2 test for independence as discussed in class and in section 4 of Chapter 28 in the book. Adding in the row and column totals for the given data, we get:

Married	Widowed, etc.	Never married	Totals
790	98	209	1097
56	11	27	94
21	7	13	41
867	116	249	1232

The first thing we need is the corresponding table of expected values under the assumption that marital status and work status *are independent*. For instance, if that is so, then the number of men in the Married/Employed category should be in the same proportion of the total Married category as the Employed category is in the whole population:

$$\frac{1097}{1232} \times 867 \doteq 772.$$

The other entries are computed similarly:

Married	Widowed, etc.	Never married	Totals
772	103.3	221.7	1097
66.2	8.9	19	94
28.9	3.9	8.3	41
867	116	249	1232

(Note the totals are off here because of rounding in a few cases, but that is not a big problem!) Now we compute the *total* χ^2-statistic comparing the two tables (9 terms altogether; some omitted for space reasons):

$$\chi^2 = \frac{(790-772)^2}{772} + \cdots + \frac{(13-8.3)^2}{8.3} \doteq 14.14.$$

The table was 3×3, so the total number of degrees of freedom if we fix the row and column totals as here is $(3-1)\times(3-1) = 4$. For 4 degrees of freedom, from the χ^2 table, we see that $p < .01$. This is strong evidence that the differences in employment status for the different marital status categories *are not due to chance*.

Chapter 28/9

(a) If the data does not come from a random sample, then no conclusions can be reasonably drawn and a χ^2 test may be meaningless.

(b) If the data does come from a random sample, then we can do a χ^2 test to see whether there is a difference (Question A). This comes down to asking whether age and marital status are *independent*. However, this type of test does not address why such a difference might have come to be (Question B). The details are similar to what we did in Chapter 28/3 above: Adding in the row and column totals for the given data, we get:

Age 20-24	Age 25-29	Totals
46	21	67
17	32	49
1	6	7
64	59	123

The expected values are computed by the same process used in the other problem:

Age 20-24	Age 25-29	Totals
34.9	32.1	67
25.5	23.5	49
3.6	3.4	7
64	59	123

Now we compute the *total* χ^2-statistic comparing the two tables (6 terms altogether):

$$\chi^2 \doteq 17.1.$$

The table was 3×2, so the total number of degrees of freedom if we fix the row and column totals as here is $(3-1) \times (2-1) = 2$. For 2 degrees of freedom, from the χ^2 table, we see that $p < .01$. This is strong evidence that the differences in marital status for the age categories *are not due to chance.*

Chapter 25/2. The genetic model would be that the smooth trait is dominant and the wrinkled trait is recessive. For the second generation hybrids, then, the inheritance of the smooth/wrinkled traits is like a draw from the box $[1, 1, 1, 0]$ (where $1 =$ smooth and $0 =$ wrinkled). That is, 75% of the offspring seeds should be smooth, and 25% should be wrinkled. The most natural way to test whether Mendel's data is close to what is expected is probably to do a χ^2 test for goodness of fit as discussed in Chapter 28. (An equivalent z-test is also possible.) Here for the 7324 plants, the expected numbers of smooth and wrinkled seed plants are:

$$7324 \times .75 = 5493 \quad \text{and} \quad 7325 \times .25 = 1831.$$

The χ^2 statistic is

$$\chi^2 = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1831} \doteq .2629.$$

For 1 degree of freedom, this gives $.50 < p < .70$. The variations from the expected values are most likely due to chance.

An alternative way to answer this would be to consider the box model above. For a z-test for how significant the number 5474 of "smoothies" is, we would do

$$z = \frac{5474 - 5493}{\sqrt{7324}\sqrt{(.75)(.25)}} \doteq -0.51$$

The p-value for this test would be estimated from the area under the standard normal curve to the left of $-.5$ and to the right of $+.5$. This would be

$$100\% - A(.5) = 100\% - 38.29\% = 61.71\%.$$

Hence $p = .6171$, which is in the range above. (This is similar to the example we discussed in class. The z-statistic, when squared, will yield the χ^2-statistic computed by the other method.)

Chapter 25/3. For the given information, the chances of getting an intermediate-flowering cross is like the chance of drawing a 1 from the box $[0, 1, 1, 0]$. Hence for the 2500 plants, we would have $EV = 1250$, and we have SE for sum of 1's is $SE = \sqrt{2500}\sqrt{(.5)(.5)} = 25$. The chance of 1300 or more intermediate-flowering plants is the chance that a standard normal has a value bigger than or equal to

$$\frac{\text{observed} - \text{expected}}{SE} = \frac{1300 - 1250}{25} = 2.0$$

This is a bit less than 2.5%.