USP 634
Spring 2017
Assignment 3: Hypothesis testing (chi-square; ANOVA)

1. A planner suspects her board is biased against affordable housing projects for fiscal zoning reasons. In recent years, the board has been unwilling to approve zoning variances for residential projects yet has been fairly lenient in approving variances for commercial and industrial projects. To test her hunch, she compiles data on board actions on zoning variance requests over the past three years, producing the contingency table shown below. Use the 5-step hypothesis testing process to test (at the 1% significance level) whether variance approval is independent of type of land use. (10 pts)

Action on Zoning Variance Request	Land Use Category		
	Residential	Commercial	Industrial
Approve	20	150	130
Deny	90	25	27

The null hypothesis here is that variance decisions are unrelated to the land use type of the request. The alternative hypothesis is that they are; specifically, that the board is more likely to award variances to commercial and industrial uses than to residential uses. For this question the appropriate test is a chi-squared test of independence, the null hypothesis of which is that land use and zoning variance actions are independent, and that any variation that we observe is simply a function of random chance and not a systematic choice process.

The chi-squared statistic is distributed with a chi-squared distribution with degrees of freedom equal to (r-1)*(c-1) where r=the number of rows and c=the number of columns. Since this is a 2x3 table the number of degrees of freedom is (2-1)*(3-1)=2.

The chi-squared test is calculated as follows:

$$u = \sum_{i=1}^{n} \frac{(f_o - f_e)^2}{f_e}$$

where f_o = cell frequences in the bivariate table

f_e = expected cell frequencies if variables are independent

n = number of cells in the table

These calculations are summarized below:

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$
20	74.66	-54.66	2987.78	40.02
150	118.78	31.22	974.8	8.21
130	106.56	23.44	549.38	5.16
90	35.34	54.66	2987.78	84.55
25	56.22	-31.22	974.8	17.34

27	50.44	-23.44	549.38	10.89
			Chi2	**166.16**

The chi-squared value of 166.16 is quite large, exceeding the critical value of 9.210 for a 99-percent confidence level (1 percent critical region) and 2 degrees of freedom. This suggests that the null hypothesis can be rejected and that the board is favoring commercial and industrial uses over residential uses in its variance granting process. [Answers that did not provide specific information about the meaning / implication of the result did not get full credit.]

2. A stratified random sample of people was asked about their race/ethnicity and their attitude to multifamily housing development in their neighborhoods. The findings were as follows:

	African-American	Hispanic	Caucasian
Approve MFH development	35	20	50
Disapprove MFH development	19	8	75

(A) Without carrying out a formal test of association, does race/ethnicity appear to be associated with one's attitude toward multifamily housing development? Why? (5 pts)

About 65 percent of African Americans support multifamily housing, and about 71 percent of Hispanics. But only about 40 percent of Caucasians do. These are pretty big differences. The column percentages change, so the variables seem to be associated. It's a relatively strong association: decreasing the level of support from 65 percent (among African-Americans, the second largest group) to 40 percent (among Caucasians, the largest group) is a reduction of a third. The "maximum difference" is between Hispanics and Caucasians is 31.4 percent which is a "strong" relationship.

(B) Carry out a test of *statistical significance* at the 95% level using Chi-squared and explain the results. (10 pts)

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$
35	27.4	7.6	57.76	2.11
20	14.2	5.8	33.64	2.37
50	63.4	-13.4	179.56	2.83
19	26.6	-7.6	57.76	2.17
8	13.8	-5.8	33.64	2.44
75	61.6	13.4	179.56	2.91
			Chi2	**14.83**

There are two degrees of freedom -- recall that degrees of freedom = (r-1)*(c-1) = (2-1)*(3-1)=2. The critical value for a chi-squared distribution with two degrees of freedom and critical region of 0.05 is 5.991. Since our chi-squared result of 14.83 is higher than the critical value we can reject the null hypothesis and conclude the observed difference are statistically significant.

(C) Give two different stories explaining the association--one that is causal, and one that not causal. (5 pts)

Two possible causal stories: (a) African Americans and Hispanics are culturally more tolerant of high density housing because of historically higher density living conditions and larger households; (b) African Americans and Hispanics are more likely to occupy multifamily housing, and while both African Americans and Hispanics tend to be tolerant of living near other African Americans or Hispanics, Caucasians are more likely to be racially biased and therefore do not support multifamily housing.

One possible non-causal story: Lower income households tend to prefer multifamily housing because it is cheaper. African Americans and Hispanics are more likely to be of lower income. Income causes the apparent association.

3. (20 pts).

4. A random survey of 32 suburban office complexes in Chico produced the following data. The dependent variable (e.g., the ratio-scale values) represents FAR, or floor-to-area ratio (building area divided by land area). The independent variable expresses four classes of office space: AAA; AA; A; and B. Class AAA normally features 20,000+ sq. ft. floor plan, fiber optics, and ultra-modern designs. At the other extreme, class B represents older buildings in less desirable locations and with fewer modern amenities.

FAR by Class of Office Space

AAA	AA	A	B
2.5	2.2	1.3	0.3
4.0	3.1	1.9	0.8
5.1	2.8	2.2	0.3
4.8	2.1	1.8	1.0
2.0	2.2	1.3	0.7
4.7	3.0	2.2	1.1
3.9	2.8	1.5	0.3
3.7	2.3	1.7	0.2

Carry out the following tests and evaluations. To get full credit you must show all your work (15 points):

(A) Use the 5-step hypothesis testing process to see if there's a significant difference in suburban office densities among classes of buildings. Test for independence at the 5% significance level.

Step 1: Assumptions:

- independent random samples
- level of measurement is interval-ratio
- the variable FAR is normally distributed
- the population variances are equal among subgroups defined by office space class

Note: Chi-squared is not an appropriate test here. In this case chi-squared requires reducing an interval-ratio variable to an ordinal variable, basically throwing away information.

STEP 2: STATE HYPOTHESES.
The null hypothesis is that there is no difference in suburban office density among classes of buildings:
H0: mean_FAR$_{AAA}$ = mean_FAR$_{AA}$ = mean_FAR$_{A}$ = mean_FAR$_{B}$
Ha: At least one class of office space has higher density than one other class of office space

STEP 3: SELECT SAMPLING DISTRIBUTION AND CRITICAL REGION
The appropriate sampling distribution for analysis of variance is the **F distribution** (using ANOVA to test for significance).

Degrees of freedom within are **32-4=28**. Degrees of freedom between=**4-1=3**. Assuming an alpha level of 0.05, the critical value is **2.95**.

STEP 4: COMPUTE THE TEST STATISTIC

$$SST = \sum (X_i - \bar{X})^2 = 56.4$$

$$SSB = \sum_{k=B,A,AA,AAA} N_k(\bar{X}_k - \bar{X})^2 = 8*(3.84-2.18)^2 + 8*(2.56-2.18)^2 + 8*(1.74-2.18)^2 + 8*(0.59-2.18)^2$$

$$= 22 + 1.2 + 1.6 + 20.2 = 45$$

$$SSW = SST - SSB = 11.4$$

$$\text{mean square between} = 45/(k-1) = 45.7/3 = 15$$

$$\text{mean square within} = 11.4/(N-k) = 11.4/28 = 0.41$$

$$F = msqb/msqw = 15/0.41 = 36.8$$

STEP 5: EVALUATE TEST STATISTIC / MAKE DECISION
The obtained value for F of 41.67 is far higher than the critical value of 2.95. We can reject the null hypothesis that average FAR does not vary among classes of office space.

This was what the class missed on #7 – interpreting the nature of the difference, using the data:

Based on this test result along with an inspection of the data, it does appear that higher classes of office space have higher density.

To help interpret ANOVA or chi-squared tests, you should always inspect conditional means, column percentages, etc to help determine WHAT difference is statistically significant! Not merely that there exists some difference that is statistically significant but not say what it is!

(B) Using the words “increase”, “decrease”, or “no effect”, state for each of the following what effect the changes described will have on the probability of finding a significant difference among groups when using ANOVA (10 pts):

(i) The standard deviation within each group is increased (without changing their means).
Decrease probability, because the calculated F statistic gets smaller when within-group variance gets larger.

(ii) The differences in group means are widened (while the variance within each category remains the same).
Increase probability, because the calculated F statistic gets larger when the difference between group means and the unconditional mean gets larger.

(iii) The group means and variations (between and within) of your samples remain the same, but your sample size doubles.
Increase probability, because the calculated F statistic gets larger when N gets larger.

(iv) You change the significance level from .05 to .01.
Decrease probability, because the critical F statistic associated with smaller regions is larger. There is in this case no effect from such a change, but the probability has increased generically speaking.

5. (20 pts).