Inferences for Counts in Tables
(Contingency Tables)

Readings: DVB Ch 26 p638- 653

- Often we are required to analyze counts in tables.
- We have seen that we can compare 2 probabilities using a Z test, and summarize the results using relative risks or odds ratios.
- Here we consider larger tables.

Example 1: A trial compared the effects of para-amino-salicylic acid (PAS) and streptomycin in the treatment of pulmonary tuberculosis, with the results shown below.

	Sputum			
Treatment	Pos. smear	Neg. smear, pos. culture	Neg. smear, neg. culture	Total
PAS	56	30	13	99
Strept.	46	18	20	84
Strept. + PAS	37	18	35	90
Total	139	66	68	273

- The treatment (row) totals are fixed by the study design.
- Our goal is to assess whether the distribution of sputum results is the same in the three rows, i.e. whether the distributions are homogeneous.
- The hypotheses for a **test of homogeneity** are
  H_0 : the distribution of results is the same in the three rows
  H_a : the distribution of results is not the same in the three rows
- Equivalently
  H_0 : the distributions of sputum results are homogeneous
  H_a : the distributions of sputum results are not homogeneous.

Example 2: Investigators wished to determine whether women infected with HIV were also likely to be infected by HPV. Results obtained for 96 women are shown below.

	HIV status			
HPV status	Sero+ sympt	Sero+ asympt	Sero-	Total
Pos	23	4	10	37
Neg	10	14	35	59
Total	33	18	45	96

- This is a cross-sectional study, and only the total number of subjects is fixed by the study design.
- We are interested in whether there is an association between HIV and HPV status, or whether they are independent.
- The hypotheses for a test of independence are
  H_0 : there is no association between HIV and HPV status
  H_a : there is an association between HIV and HPV status.
- It is equivalent to ask whether the distribution of HIV status is the same regardless of the HPV status, or whether the distribution of HPV status is equivalent for each HIV status.
- So the hypotheses are similar in the two examples, despite the difference in study design, and in fact the tests are done in exactly the same way!

The χ^2 test

- The χ^2 test statistic compares the observed counts to those which are expected if the null hypothesis is true

$$X^2 = \sum\sum \frac{(obs - exp)^2}{exp}.$$

- The sums are over all cells in the table.
- The expected counts are given by

$$exp = \frac{row\ sum \times column\ sum}{overall\ sum}.$$

- If the counts are in agreement with the null hypothesis, X^2 will be small.
- So large values give evidence against the null hypothesis.

The χ^2 distribution and P value

- If the null hypothesis is correct, X^2 approximately has a χ^2 distribution, with degrees of freedom $(r-1)(c-1)$.
  - r is the number of rows
  - c is the number of columns
- For this approximation to be valid, all the expected counts should be at least 5.
- If some expected counts are less than 5, then a version of Fisher's exact test can be used.
- The P value is the probability in the right tail of the χ^2 distribution beyond the observed value.
- Using tables, we can usually only obtain bounds on the P value.

Solving example 1 (TB): The expected counts are

	Sputum			
Treatment	Pos. smear	Neg. smear, pos. culture	Neg. smear, neg. culture	Total
PAS	50.41	23.93	24.66	99
Strept.	42.77	20.31	20.92	84
Strept. + PAS	45.82	21.76	22.42	90
Total	139	66	68	273

- For example, in the first cell

$$exp = \frac{99(139)}{273} = 50.41$$

- Note
  1. We do not round these values to integers!!
  2. The expected counts add up to the observed totals for the rows and columns.

- It can be helpful to show the contributions to the overall test statistic, $(obs - exp)^2/exp$, as these can reveal where the departures occur.

	Sputum			
Treatment	Pos. smear	Neg. smear, pos. culture	Neg. smear, neg. culture	Total
PAS	.62	1.54	5.51	
Strept.	.24	.26	.04	
Strept. + PAS	1.70	.65	7.06	
Total				17.63

- The biggest contributions occur in the last column of the first and third rows.

- There is smaller observed count for Negative smear, negative culture in the PAS group (13 vs 24.66), and a larger observed count in the Strept. + PAS group (35 vs 22.42).

- There are (3-1)(3-1) = 4 degrees of freedom.
- Comparing the test statistic $X^2 = 17.63$ to the table, we find that 17.63 exceeds the largest value 14.86, so $P < .005$.
- We therefore have very strong evidence against the null hypothesis of no difference in the distributions of sputum results among the treatment groups.

Solving example 2: HIV/HPV status

- The expected counts are

	HIV status			
HPV status	Sero+ sympt	Sero+ asympt	sero-	Total
Pos				37
Neg				59
Total	33	18	45	96

- The contributions to X^2 are

	HIV status			
HPV status	Sero+ sympt	Sero+ asympt	sero-	Total
Pos				
Neg				
Total				

- The test statistic is $X^2 =$ on df.
- The P value is
- We conclude

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	C1	C2	C3	Total
1	23	4	10	37
	12.72	6.94	17.34	
	8.311	1.244	3.110	
2	10	14	35	59
	20.28	11.06	27.66	
	5.212	0.780	1.950	
Total	33	18	45	96

Chi-Sq = 20.606, DF = 2, P-Value = 0.000