Chi-square (χ^2) Tests

- **Common Uses of the χ^2-test.**

  1. Testing Goodness-of-fit.

  2. Testing Equality of Several Proportions.

  3. Homogeneity Test.

  4. Testing Independence.

- **Testing Goodness-of-fit of Data from Multinomial Experiment.**

  Properties of the Multinomial Experiment.

  1. The experiment consists of n identical trials.

  2. There are k possible outcomes to each trial. These possible outcomes are sometimes called *classes* or *categories*.

  3. The probabilities of the k outcomes, denoted by $p_1, p_2, \ldots, p_k$, remain the same from trial to trial, and $p_1 + p_2 + \cdots + p_k = 1$.

  4. The trials are independent.

  5. The random variables of interest are the *cell counts*, $n_1, n_2, \ldots, n_k$, of the number of observations that fall in each of the k categories. (note: $n_1 + n_2 + \cdots + n_k = n$).

	Cat 1	Cat 2	$\cdots$	Cat k
Observed	n_1	n_2	$\cdots$	n_k
Expected	$n * p_1$	$n * p_2$	$\cdots$	$n * p_k$

- **Chi-Square Statistic.** The *chi-square statistic* is a measure of how much the observed cell counts diverge from the expected cell counts. The formula for the statistic is

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{k} \frac{(observed\ count - expected\ count)^2}{expected\ count} && (1)\\
&= \sum_{i=1}^{k} \frac{(O-E)^2}{E} && (2)\\
&= \sum_{i=1}^{k} \frac{(n_i - n * p_i)^2}{n * p_i} \sim \chi^2_{k-1}. && (3)
\end{aligned}
$$

  This X^2 statistic follows **approximately** the χ^2 distribution with $k-1$ degrees of freedom.

  Note: *The chi-square approximation is adequate for practical use when the average expected cell is 5 or greater and all individual expected counts are at least 1, except in a 2×2 table where all four expected counts should be at least 5.*

- **Null Hypothesis and Alternative Hypothesis.**

  1. Null Hypothesis: $H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots, p_k = p_{0k}$

  2. Alternative Hypothesis: H_1 : The null hypothesis is not true.

- **Example:**

1. *Nature* (Sept. 1993) reported on a study of animal and plant species "hotspots" in Great Britain. A hotspot is defined as a 10-km^2 area that is species-rich, that is, is heavily populated by the species of interest. Analogously, a coldspot is a 10-km^2 area that is species-poor. The table below gives the number of butterfly hotspots and the number of butterfly coldspots in a sample of 2,588 10-km^2 areas. In theory, 5% should be butterfly hotspots and 5% coldspots, while the remaining areas (90%) are neutral. Test the theory using $\alpha = 0.01$.

	hotspots	coldspots	Neutral Areas	Total
Observed	123	147	2318	2588
Expected				

2. Car Crashes and Age Brackets. Among drivers who have had a car crash in the last year, 88 are randomly selected and categorized by age, with the results listed in the accompanying table. If all ages have the same crash rate, we would expect (because of the age distribution of licensed drivers) the given categories to have 16%, 44%, 27%, and 13% of the subjects, respectively. At the 0.05 level of significance, test the claim that the distribution of crashes conforms to the distribution of ages. Does any age group appear to have a disproportionate number of crashes?

Age	Under 25	25 - 44	45 - 64	Over 64
No. of Crashes	36	21	12	19

3. Checking Normality. The table below shows the number of values in a sample of size $n = 200$ observations falling in each category. At the 0.05 level of significance, test whether the sample can be reasonably assumed to have come from the Standard Normal distribution.

$z-$values	$z < -1$	$-1 \leq z < -.5$	$-.5 \leq z < 0$	$0 \leq z < .5$	$.5 \leq z < 1$	$z > 1$
No. of Obs.	30	38	43	38	20	31

- Testing Equality of Several Proportions.
  1. Null Hypothesis: $H_0 : p_1 = p_2 = \cdots = p_k$
  2. Alternative Hypothesis: H_1 : Not all are equal.

- Examples.
  1. **No Smoking.** The accompanying table summarizes successes and failures when subjects used different methods in trying to stop smoking. The determination of smoking or not smoking was made five months after the treatment was begun, and the data are based on results from the Centers for Disease Control and Prevention. Use a 0.05 level of significance to test the claim that success is independent of the method used. If someone wants to stop smoking, does the choice of the method make a difference?

	Nicotine Gum	Nicotine Patch	Nicotine Inhaler
Smoking	191	263	95
No smoking	59	57	27

  2. **Survey Refusals and Age Bracket.** A study of people who refused to answer survey questions provided the randomly selected sample data shown in the table. At the 0.01 significance level, test the claim that the cooperation of the subject (response or refusal) is independent of the age category. Does any particular age group appear to be particularly uncooperative?

Age	18-21	22-29	30-39	40-49	50-59	60 and over
Responded	73	255	245	136	138	202
Refused	11	20	33	16	27	49

- Testing for Homogeneity (In a Two-Way Contingency Tables).

  1. Null Hypothesis: $H_0 : p_{1j} = p_{2j} = \cdots = p_{kj},\ j = 1, 2, \ldots, J$

  2. Alternative Hypothesis: $H_1 : H_0$ is not true.

  3. Test Statistic: The *chi-square statistic* X^2 is a measure of how much the observed cell counts diverge from the expected cell counts. The formula for the statistic is very similar to the one we have before.

$$
\begin{aligned}
X^2 &= \sum_{\text{all cells}} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} && (4)\\
&= \sum_{i=1}^{k}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{J} \frac{(n_{ij} - r_i * \hat{p}_j)^2}{r_i * \hat{p}_j} && (5)\\
&= \sum_{i=1}^{k}\sum_{j=1}^{J} \frac{(n_{ij} - r_i * c_j/n)^2}{r_i * c_j/n} \sim \chi^2_{(k-1)(J-1)}. && (6)
\end{aligned}
$$

  where, r_i and c_j are the total of the ith row and jth column, respectively. This X^2 statistic follows **approximately** the χ^2 distribution with $(k-1)(J-1)$ degrees of freedom.

  Note: *The chi-square approximation is adequate for practical use when the average expected cell is 5 or greater and all individual expected counts are at least 1, except in a* 2×2 *table where all four expected counts should be at least 5.*

- Example.

  1. A company packages a particular product in cans of three different sizes, each one using a different production line. Most cans conform to specifications, but a quality control engineer has identified the following reasons for nonconformance: Blemish on can, crack in can, improper pull tab location, pull tab missing, and other. A sample of nonconforming units is selected from each of the three lines, and each unit is categorized according to reason for nonconformity, resulting in the following contingency table data:

Production	Blemish	Crack	Location	Missing	Other	Total
Line 1	34	65	17	21	13	150
Line 2	23	52	25	19	6	125
Line 3	32	28	16	14	10	100
Total	89	145	58	54	29	375

- Testing for Independence (In a Two-Way Contingency Tables).
  1. Null Hypothesis: Factor A is independent of factor B.
  2. Alternative Hypothesis: Factors A and B are not independent.
  3. Test Statistic: The same *chi-square statistic* X^2.

- Example.
  1. **Background Music.** Market researchers know that background music can influence the mood and purchasing behavior of customers. One study in a supermarket in Northern Ireland compared three treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. Here is the two-way table that summarizes the data:

Wine	French music	Italian music	No music
French	39	30	30
Italian	1	19	11
Other	35	35	43

Homework problems: Worth 10 points. Due Wednesday, 4/22/08.

Section 13.2: pp. 791-794; # 3, 7, 8, 11, 12.
Section 13.3: pp. 805-811; # 21, 22, 24, 26.
Supplement: pp. 813-817 # 40, 41, 42, 45.