Chapter 14: Inference for Categorical Data

- In some studies, the data gathered are _____ rather than _____.
- Each individual or experimental unit is classified into a _____.
- We can summarize such data by finding the _____ of how many measurements fall into each category (cell).
- When there are two categories, such data can often be modeled as a _____ _____.
- When there are more than two categories, such data can often be modeled as a _____ _____.
- Examples: ① Students in a class could be classified into grade categories (A, B, C, D, F).
② Music recordings could be classified into one of several genres.

- Recall the conditions for a _____ experiment:

(1) There are n identical and independent trials, each having as its outcome one of k distinct categories.

(2) The probability of a trial being in category i is p_i $(i=1,2,\ldots,k)$ where

and the set of p_i's is the same for each trial.

(3) The multinomial r.v. is the set of counts of the number of trials that fall in each category.

- Note

A Test About Multinomial Probabilities

- Suppose we rolled $n=300$ balanced six-sided dice. How many "1" results do we expect?
- That is, what is $E(n_1)$?

- Marginally, n_1 has a _____ distribution with parameters and
- So $E(n_1) =$
- Similarly, each of $n_2, n_3, \ldots, n_6$ has a marginal _____ distribution, so $E(n_i) =$

- Now, suppose we might not know the dice are balanced (but maybe we hypothesize that they are).
- To test whether a set of hypothesized category probabilities $p_1, p_2, \ldots, p_k$ are correct, we could compare the observed counts (the) to their expected values if the probabilities were true, i.e., look at

- Karl Pearson proposed using this statistic:

Theorem: When n is large, X^2 has a

Proof: (We will prove this only in the $k=2$ case.)

Since $n_1 \sim$

- The proof for general k can be shown via mathematical induction, but there are subtle technicalities related to independence of terms in the summation.

χ^2 Goodness-of-Fit Test

- Let _____ be the hypothesized cell probabilities.
- Consider testing

H_0:

vs H_a:

- If H_0 is false, we would expect $[n_i - E(n_i)]^2$ to be _____ and the test statistic

to be _____.

- So we reject H_0 if _____
- In general, the degrees of freedom for this χ^2 distribution are the number of cells k minus one d.f. for each <u>independent</u> <u>linear</u> <u>restriction</u> on the cell probabilities.

- For example, we have the restriction

so already we have degrees of freedom in this basic goodness-of-fit test.

- Furthermore, if the probabilities in H_0 are estimated from a sample rather than being completely specified, then the degrees of freedom are

- Furthermore, if unknown parameters are to be estimated, they must be estimated by the method of

_____ _____ .

- How large should n be for the χ^2 test to be valid?

Rule of Thumb: Each expected count $E(n_i) = np_i$ should be

- This is overly onerous: W.S. Cochran showed the approximation can be good even if some $E(n_i)$

- If any expected counts are too small to use the χ^2 test, one remedy is to combine certain categories and do the test with a reduced number of cells (if this makes sense).

Example 1: The manufacturer of M+M's states that for plain M+M's, 13% should be brown, 14% yellow, 13% red, 24% blue, 20% orange, and 16% green. In a randomly selected bag of M+M's, we find the following color counts:

Brown	Yellow	Red	Blue	Orange	Green
61	64	54	61	96	64

- Test the manufacturer's claim, using $\alpha = 0.05$.

$n = 61 + 64 + \cdots + 96 + 64 =$

Expected counts:

Example 2(a): Consider counts of the number of V-2 rockets landing in each of 576 regions on a 24×24 grid map of London. Do the counts follow a Poisson distribution with mean 1? Use $\alpha = 0.05$.

# of rockets	0	1	2	3	4
Count of regions	229	211	93	35	8

Example 2(b): Same data, but suppose the question was simply whether the data follow a Poisson distribution.

- Now we must

Note: We have used a goodness-of-fit test to test whether a population followed a Poisson distribution.

- We could use a goodness-of-fit test similarly to test whether data came from a continuous distribution (like a normal), if our data were continuous.
- We would have to break the support of the distribution into categories and find the probability under H_0 of an observation falling into each category (and thus the expected counts of data falling in each category).
- These expected counts would be compared to the observed category counts using

- If the parameters of the distribution were not specified, they would be estimated with MLE's, and we would subtract