Notes for Thursday February 19, 1998

A classification of data and analysis types

So far in 401 and 402 we have concentrated on continuous-response data of one type or another. We will now turn to analyzing data in which the responses are continuous. The following table shows how this fits into the general scheme of statistics, and in particular where logistic regression (which we will talk about) and regression/ANOVA (which we have talked about) fit in.

	Explanatory Variables		
Response	Categorical	Continuous	Both
Categorical	Contingency Tables	Logistic Regression	Loglinear Model
Continuous	ANOVA/MANOVA	Linear Regression	ANCOVA
Both	?/Hard	?/Hard	?/Hard

Contingency Tables, Loglinear Models, and Logistic Regression

You know most of the basics for the kinds of analysis in the second line of the table above—continuous responses. We will now turn to the analysis of data in which the response variable is discrete (or categorical). There are basically two kinds of categorical data to think about: *counts'*, and *binary responses*. There are actually other kinds of discrete responses than binary ones, but once you have an intuition for binary responses the others come more easily.

Typical count data comes in the form of a contingency table, which is a cross-classification of factors, with counts of observed outcomes in every cell of the table. This is one of two useful ways to view count data, illustrated below.

		Lung Cancer?	
		Yes	No
Smoking?	Yes	360	120
	No	95	105

Contingency Table

Response	*Smoking*	*Lung Cancer*
360	1	1
95	0	1
120	1	0
105	0	0

Response and Predictors

Some fictional data to illustrate counts

The *Contingency Table* display of data is especially good for

- Computing and comparing proportions
- Thinking about independence and patterns of dependence
- Computing conditional probabilities and odds
- Sampling schemes and sampling distributions

The *Response and Predictors* display is especially good for

- Irregular tables
- Unusual dependence structures that are not readily related to conditional probabilities given a factor level
- Modeling analogous to `lm()` and `aov()`
- Generalization to logistic regression and other generalized linear models (`glm()`).

We will begin by thinking about count data in contingency tables, but before long we will move to the other way of arranging the data, since it is much more flexible.

Small Contingency Tables

Let's think about the fictional table

		Lung Cancer?		
		Yes	No	Total
Smoking?	Yes	360	120	480
	No	95	105	200
	Total	455	225	680

a bit more. A reasonable thing to think about is what sort of a statistical model might be appropriate to describe these outcomes. There are several possibilities:

- There are 680 total persons in this study. If we decided in advance to go out and sample 680 people at random (in some sort of survey, say) and note their smoking and cancer status, we might get a table like this. The appropriate distribution would be the *multinomial* distribution,

  $$(x_{11}, x_{12}, x_{21}, x_{22}) \sim Multi(680; p_{11}, p_{12}, p_{21}, p_{22})$$

  (where the counts are indexed like cells in a matrix) which means that

  $$P[seeing\ (x_{11}, x_{12}, x_{21}, x_{22})] = \left(\begin{array}{c} n \\ x_{11}, x_{12}, x_{21}, x_{22} \end{array} \right) p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} p_{22}^{x_{22}}$$

  where $n = 680$.

  It's a good 225/226 exercise to show that the *maximum likelihood estimate* of p_{ij} is (to no one's surprise)

  $$\hat{p}_{ij} = x_{ij}/n$$

  and we know from elementary statistics that since each individual count in a multinomial is itself a binomial, then

  $$\begin{aligned} \mathrm{Var}\,(x_{ij}) &= x_{ij} p_{ij}(1 - p_{ij}), \text{ so that} \\ \widehat{\mathrm{Var}}\,(\hat{p}_{ij}) &= \hat{p}_{ij}(1 - \hat{p}_{ij})/x_{ij} \end{aligned}$$

- A second possibility is that we simply sampled people for a fixed amount of time (say) rather than to a fixed sample size. Events per unit time are often modeled with a *Poisson distribution*, and here we have four different kinds of events, so we might want to model them with four different Poisson distributions,

  $$P[seeing\ (x_{11}, x_{12}, x_{21}, x_{22})] = \frac{m_{11}^{x_{11}}}{x_{11}!} e^{-m_{11}} \frac{m_{12}^{x_{12}}}{x_{12}!} e^{-m_{12}} \frac{m_{21}^{x_{21}}}{x_{21}!} e^{-m_{21}} \frac{m_{22}^{x_{22}}}{x_{22}!} e^{-m_{22}}$$

  where m_{ij} is the mean (expected) count in cell ij. Clearly (exercise!!) the MLE of m_{ij} is $\hat{m}_{ij} = x_{ij}$, and the MLE of p_{ij} is still

  $$\hat{p}_{ij} = x_{ij}/n$$

  the same as before. The main difference between multinomial sampling and Poisson sampling is that in multinomal the grand total n is fixed in advance, and in Poisson it is also a random variable. It is a *fact* that as long as we don't want to make inferences about n, it doesn't matter whether we use multinomial or Poisson sampling. Mathematically,

$$P_{Pois}[seeing\ (x_{11}, x_{12}, x_{21}, x_{22}) \mid n = 680] \ = \ P_{Multi}[seeing\ (x_{11}, x_{12}, x_{21}, x_{22})]$$

Figuring out the variance is a little harder. As with all Poisson's Var $(x_{ij}) = m_{ij}$, but $p_{ij} = x_{ij}/(x_{11} + x_{12} + x_{21} + x_{22}$. After some work, one can show that

$$\widehat{\text{Var}}\,(\hat{p}_{ij}) \ = \ \frac{x_{ij}}{n}(1 - \frac{x_{ij}}{n})/n$$

which is exactly equivalent to the variance under multinomial sampling (this is part of the equivalence of Poisson and Multinomial sampling mentioned above.

- We might have gone out and found 480 *smokers* and 200 *nonsmokers* and followed them through life to see if they ever contracted cancer. This is an example of a *prospective study*. In this case, the row totals $n_{1.} = 480$ and $n_{2.} = 200$ are fixed, and the counts x_{11} and x_{21} (say) are independent binomials:

$$\begin{aligned} P[seeing\ (x_{11}, x_{12}, x_{21}, x_{22})] \ &= \ P[seeing\ (x_{11}, x_{21}) | n_{1.}, n_{2.}] \\ = \ &\binom{x_{11}}{n_{1.}} p_{11}^{x_{11}}(1 - p_{11})^{n_{1.} - x_{11}} \binom{x_{21}}{n_{2.}} p_{21}^{x_{21}}(1 - p_{21})^{n_{2.} - x_{21}} \end{aligned}$$

Note that the MLE and variance have the same form as before, but now the denominators are different:

$$\hat{p}_{ij} = x_{ij}/n_{i.}$$

and the variance will again be

$$\widehat{\text{Var}}\,(\hat{p}_{ij}) = \hat{p}_{ij}(1 - p_{ij})/n_{i.}$$

- Finally, we could have gone out and found 455 cancer patients ("cases") and 225 noncancer patients ("controls") and asked how many were smokers. This is an example of a *retrospective study*, also known as a *case control study*. Now the column totals $n_{.1} = 455$ and $n_{.2} = 225$ are fixed, and the counts x_{11} and x_{12} are independent binomials, with

$$\begin{aligned} P[seeing\ (x_{11}, x_{12}, x_{21}, x_{22})] \ &= \ P[seeing\ (x_{11}, x_{21}) | n_{1.}, n_{2.}] \\ = \ &\binom{x_{11}}{n_{.1}} p_{11}^{x_{11}}(1 - p_{11})^{n_{.1} - x_{11}} \binom{x_{21}}{n_{.2}} p_{21}^{x_{12}}(1 - p_{12})^{n_{.2} - x_{21}} \end{aligned}$$

The MLE's are now

$$\hat{p}_{ij} = x_{ij}/n_{.j}$$

with variances

$$\widehat{\mathrm{Var}}\,(\hat{p}_{ij}) = \hat{p}_{ij}(1 - p_{ij})/n_{.j}$$

Note that *prospective* and *retrospective* studies both lead to the same sampling distribution—product-binomial. You only have to figure out if it's the rows or the columns that are the independent binomial data being considered. A useful way to do this is to decide whether one of the factors is EXPLANATORY. If so, that is the factor that splits the table into two binomials.

A one-way table

The same ideas apply to tables of any "shape". Here we consider a one-way layout of counts.

We examine 3343 families with 4 or more children and find out how many boys in the first 4 children. The data are

0	1	2	3	4
183	789	1250	875	246

So far we have not considered models. What is an appropriate model for these data? If we assume (a little erroneously) that $\Pr(\text{Boy}) = 0.5$, and that births within a family are independent then we might generate the following probabilities,

0	1	2	3	4
$\frac{1}{2}^4$	$4\frac{1}{2}^4$	$6\,\frac{1}{2}^4$	$4\frac{1}{2}^4$	$\frac{1}{2}^4$
$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

which would give us an expected table of counts of

0	1	2	3	4
208.9	835.75	1253.6	837.75	208.9

and then we need to ask how well does the data fit the model.

One test of this is called Pearson's X^2 which is

$$\begin{aligned}\mathrm{X}^2 &= \sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}} \\ &= \frac{(183-208.9)^2}{208.9} + \cdots + \frac{(246-208.9)^2}{208.9} \\ &= 14.27\end{aligned}$$

Asymptotically (i.e., as the sample size gets big) this is distributed as a χ^2_4 random variable. Why 4 df? Note that $\chi^2_4(0.05) = 9.49$, and $\chi^2_4(0.01) = 13.28$.

More 2 × 2 Examples

Example 1

Linus Pauling (Nobel laureate and Vitamin C fame) reported the following *double blind* data on French skiers and their propensity to get colds.

	Cold	No Cold	
Placebo	31	109	140
Vit C (1 gm/day)	17	122	139

How was (could have) these data been collected? Two groups were selected, one was given treatment and the other the placebo. The aim was to compare the probability of getting a cold in each group. In other words we have two binomials, one for Placebo and one for Vitamin C and we wish to compare responses. Here getting a cold is the *response* variable and the treatment is the explanatory variable.

Example 2

The following data is the cross-classification of Perch Height by Perch Diameter of 164 lizards.

Height	Diameter (in)	
(feet)	$<= 4.0$	> 4.0
> 4.75	32	11
$<= 4.75$	86	35

How were these data collected? Is the Poisson or Multinomial distribution appropriate?

What sort of question are appropriate for these data?

- Is the proportion at height 1 > height 2?
- Is the proportion at diameter 1 > diameter 2?
- Is height independent of diameter?

Whether we can answer these questions depends on how the data were sampled?

In deciding what sampling model to use, it is valuable to think about which of the factors is a RESPONSE factor and which factor is an EXPLANATORY factor.

Skiers and Vitamin C: Product Binomial

We have two groups of skiers—those that were given Vitamin C and those that were not. We wish to ask the question: Is the probability of getting a cold the same in both groups?

If the data is,

	Cold	No Cold	
Placebo	31	109	140
Vit C (1 gm/day)	17	122	139

then a reasonable model for the data might be

	Cold	No Cold
Placebo	p_1	$(1 - p_1)$
Vit C (1 gm/day)	p_2	$(1 - p_2)$

and the question becomes: Is $p_1 = p_2 = p$?

A good estimate of the common p is $\frac{(31+17)}{279}$. But how do we test if this common value of p adequately describes the data?

Consider the generic table of data,

$x_{1,1}$	$x_{1,2}$	n_1
$x_{2,1}$	$x_{2,2}$	n_2

Method 1 Note that $\hat{p}_1 = \frac{x_{1,1}}{n_1}$ and $\hat{p}_2 = \frac{x_{2,1}}{n_2}$ and we wish to test if

$$p_1 - p_2 = 0$$

Note that $\mathrm{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$ and that

$$\begin{aligned} \mathrm{Var}(\hat{p}_1 - \hat{p}_2) &= \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

If we assume that $p_1 = p_2$ then

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = (\frac{1}{n_1} + \frac{1}{n_2})(1-p)p$$

If we further assume that $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean 0 and variance $\mathrm{Var}(\hat{p}_1 - \hat{p}_2)$, then we can test the hypothesis with the statistic

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})(1-\hat{p})\hat{p}}} \sim \mathrm{N}(0,1)$$

For the skiers example, this is $= 2.19$, which is quite a large number.

We had to assume approximate normality. Numerous numerical (and theoretical) studies have shown this is a quite valid assumption even when the counts in each "cell" of the table become quite small.

Method 2 An alternative test for $p_1 = p_2$ can be obtained by using Pearson's X^2,

$$\mathrm{X}^2 = \sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$$

where O are the observed values and E are the expected values under the model.

For our type of data the expected table is

$n_1\hat{p}$	$n_1(1-\hat{p})$
$n_2\hat{p}$	$n_2(1-\hat{p})$

which is equal to

24.1	115.9
23.9	115.1

and thus $\mathrm{X}^2 = 4.81 = (2.19)^2$. For this example, the X^2 statistic is the square of a $\mathrm{N}(0,1)$—so its distribution is χ^2_1. In general, X^2 is a useful statistics that is distributed as a χ^2_*. There is some uninteresting algebra which show the equivalence of the two tests.

Why is it on 1 degree of freedom? Note that the original data had only 2 degrees of freedom because the sample sizes were predetermined. We used the data to estimate one parameter, p, which left 1 degree of freedom for the test.

This test for equality of Binomials is more generally referred to as a test of HOMOGENEITY OF PROPORTIONS. *Note the importance of identifying the* RESPONSE *and* EXPLANATORY *variables.*

Lizards on Perches: Poisson or Multinomial

We have one group of lizards which are cross-classified according to the perch height and diameter. Sampling may have been either multinomial (we knew how many lizards there were) or Poisson (we didn't know). Notice we have 2 RESPONSE variables, and no EXPLANATORY variables, so appropriate questions about the data concern the interrelationships between the responses.

Height	Diameter (in)	
(feet)	<= 4.0	> 4.0
> 4.75	32	11
<= 4.75	86	35

The data and probabilities of observation are, Consider the generic table of data,

$x_{1,1}$	$x_{1,2}$	n_1
$x_{2,1}$	$x_{2,2}$	n_2

probabilities

$p_{1,1}$	$p_{1,2}$	$p_{1,+}$
$p_{2,1}$	$p_{2,2}$	$p_{2,+}$
$p_{+,1}$	$p_{+,2}$	$p_{+,+}$

We may like to ask questions like "Is Diameter independent of Height?". For independence we know that marginal probabilities need to be multiplied. If we take the marginal Height and Diameter probabilities as given then the independence hypothesis say that

$$\hat{p} = \begin{array}{|c|c|}\hline p_{+,1}p_{1,+} & p_{+,2}p_{1,+} \\ \hline p_{+,1}p_{2,+} & p_{+,2}p_{2,+} \\ \hline \end{array}$$

should be the observed probabilities under independence.

In other words, the table of fitted values under the independence model is

$$\begin{array}{|c|c|}\hline \frac{x_{+,1}x_{1,+}}{x_{+,+}} & \frac{x_{+,2}x_{1,+}}{x_{+,+}} \\ \hline \frac{x_{+,1}x_{2,+}}{x_{+,+}} & \frac{x_{+,2}x_{2,+}}{x_{+,+}} \\ \hline \end{array} = \begin{array}{|c|c|}\hline 30.9 & 12.1 \\ \hline 87.1 & 33.9 \\ \hline \end{array}$$

The value of X^2 for these data is 0.18 which should be compared to a χ^2_1—not very large.