More on 2 × 2 Tables
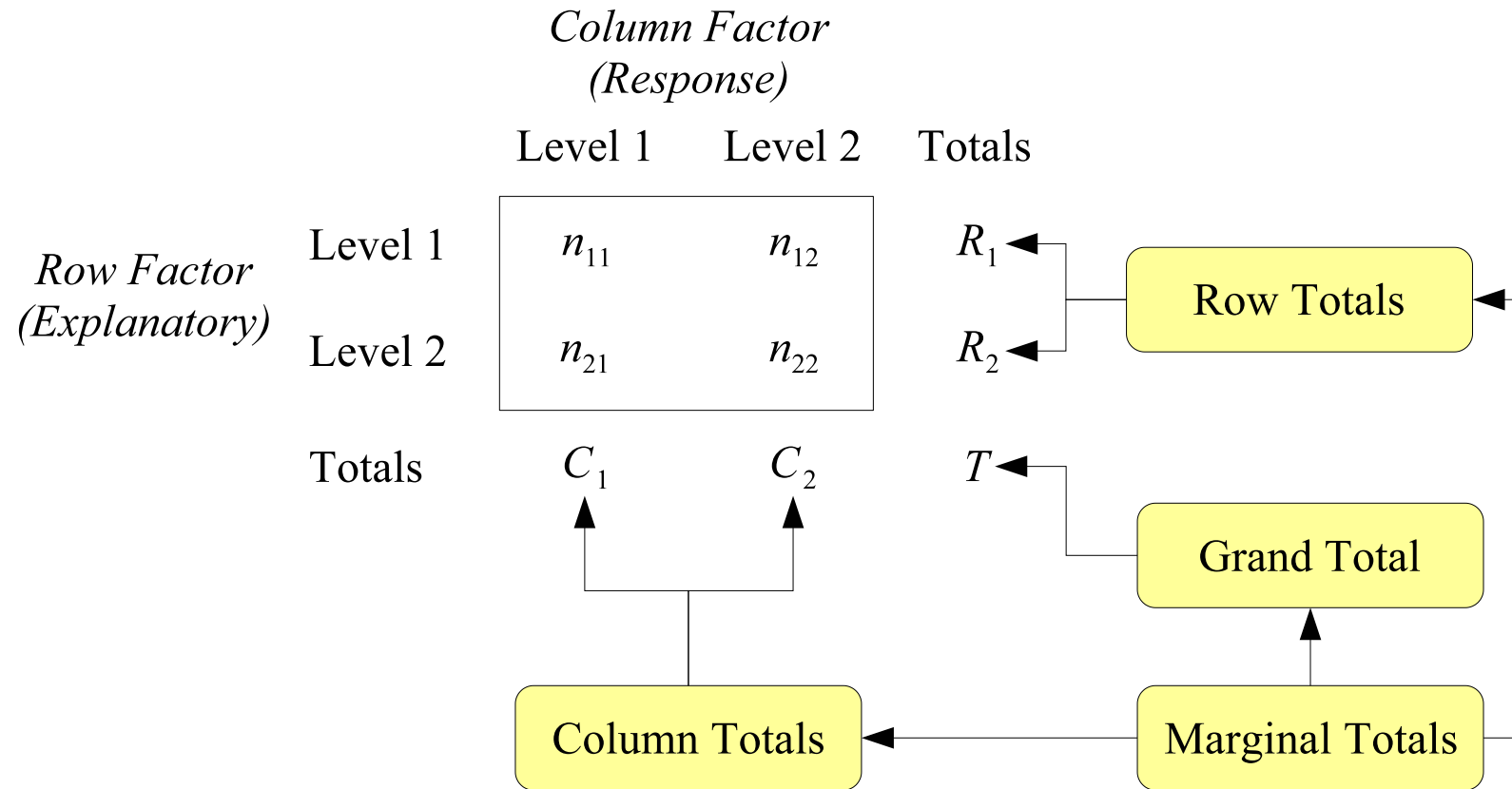
Statistics 149

Spring 2006

Sampling Schemes for 2×2 Tables

In the previous examples, the data could be put into a 2×2 tables.

Column Factor
(Response)

	Level 1	Level 2	Totals
Level 1	n_{11}	n_{12}	R_1
Level 2	n_{21}	n_{22}	R_2
Totals	C_1	C_2	T

Row Factor
(Explanatory)

Row Totals

Grand Total

Column Totals

Marginal Totals

However, the mechanisms (i.e. the sampling schemes) used to collect the data varied. They included

- Experiment (Aspirin trial to prevent strokes)

- Two independent samples (Lipid drugs and risk of Rhabdomyolysis)

- A single sample cross classified by two variables (Infant deaths in New York)

As we saw in comparing prospective to retrospective studies, the parameters that can be estimated and the type of comparisons that can be made may depend on the sampling mechanism.

There are 6 common sampling schemes that are seen that lead to a 2×2 table structure.

- **Poisson Sampling:** A sample from a single population is sampled and each member falls into one of the four cells of the 2×2 table.

  An example is the infant deaths in New York. Each child born was classified by birth weight (≤ 2500g vs > 2500g) and mortality status at one year.

  In this scheme none of the marginal totals (or grand total) are known. In the infant death example, the total number of births in New York in 1974 is a random variable.

- **Multinomial Sampling:** This scheme is related the Poisson scheme, except the grand total is a fixed quantity. As in the Poisson scheme, each subject sampled falls into one of the four cells of the table. This could be done, for the infant death example, by taking say, a random sample of 5000 from the total of 37840 births that year.

- **Prospective Product Binomial Sampling:** This scheme matches with the motivation of the last chapter. Two binomial samples, one or each level of the explanatory variable, are taken. For each group, the sample size (row total) is fixed. Then each subject/object is classified by the levels of the response variable.

- **Retrospective Product Binomial Sampling:** This is equivalent to the prospective product binomial sampling case where the sampling subpopulations are defined by the levels of the response factor (fixing the column totals) and the subjects are then classified by the explanatory variable.

- **Randomized Binomial Experiment:** In this case, the subjects are randomly allocated to the two levels of the explanatory variable, thus fixing them (row totals). Then each subject is classified by the response variable. This is effectively the same scheme as prospective product binomial sampling.

- **Hypergeometric Probability Distribution:** In this case both the row and column totals are fixed. Observations are classified by both variables, subject to both row and column constraints.

  The classic example Fisher's Tea Tasting Experiment (1935, Experiement Design, page 11) a lady is asked to sample 8 cups of tea, 4 where the milk is poured first, and 4 with tea poured first (fixing the row totals). The lady is informed that there will be 4 of each (fixing the row totals). Hypothetical data looks like

	Guess		
Truth	Milk First	Tea First	Total
Milk First	3	1	4
Tea First	1	3	4
Total	4	4	8

Sampling Scheme	Marginal Totals Fixed
Poisson	None
Multinomial	Grand Total
Product Binomial	
Prospective	Row (explanatory) Totals
Retrospective	Column (explanatory) Totals
Randomized Experiment	Row (explanatory) Totals
Hypergeometric	Row and Column Totals

Column Factor
(Response)

		Level 1	Level 2	Totals
	Level 1	n_{11}	n_{12}	R_1
Row Factor *(Explanatory)*	Level 2	n_{21}	n_{22}	R_2
	Totals	C_1	C_2	T

Probability Models Underlying Sampling Schemes

Underlying the analysis of each of these sampling schemes are probability models. The following are the usual assumptions made in each case (at least one way of describing them)

- **Prospective Product Binomial Sampling:**

  1. Explanatory level 1: $n_{11} \sim Bin(R_1, \pi_1)$
  2. Explanatory level 2: $n_{21} \sim Bin(R_2, \pi_2)$

  and n_{11} and n_{21} are independent.

	Level 1	Level 2	Total
Level 1	π_1	$1 - \pi_1$	1
Level 2	π_2	$1 - \pi_2$	1

  (Note the randomized experiment case follows the same model.)

- **Retrospective Product Binomial Sampling:**

  1. Response level 1: $n_{11} \sim Bin(C_1, \nu_1)$
  2. Response level 2: $n_{12} \sim Bin(C_2, \nu_2)$

  and n_{11} and n_{12} are independent.

	Level 1	Level 2
Level 1	ν_1	ν_2
Level 2	$1 - \nu_2$	$1 - \nu_2$
Totals	1	1

  This is effectively the same scheme. Instead we are conditioning on columns instead of rows.

  (Note in last class ν_1 and ν_2 were called p_1 and p_2. I want to use p for something else today.)

- **Multinomial Sampling:**

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim Multi(T, (p_{11}, p_{12}, p_{21}, p_{22}))$$

where $p_{11} + p_{12} + p_{21} + p_{22} = 1$

	Level 1	Level 2	Total
Level 1	p_{11}	p_{12}	p_{1+}
Level 2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

(the $+$ in the subscript indicates to add across that index).

With this type of sampling, we can learn about similar quantities as in the two product binomial sampling cases. This is due to the fact, that if $(n_{11}, n_{12}, n_{21}, n_{22})$ has the above Multinomial distribution, then

$$R_1 \sim Bin(T, \rho)$$
$$C_1 \sim Bin(T, \gamma)$$
$$n_{i1} | R_i \sim Bin(R_i, \pi_i) \qquad \text{Row } i$$
$$n_{1j} | C_j \sim Bin(C_j, \nu_j) \qquad \text{Column } j$$

where

$$\rho = p_{+1} \qquad \gamma = p_{1+}$$

$$\pi_i = \frac{p_{i1}}{p_{i1} + p_{i2}} = \frac{p_{i1}}{p_{i+}}$$

and

$$\nu_j = \frac{p_{1j}}{p_{1j} + p_{2j}} = \frac{p_{1j}}{p_{1+}}$$

- **Poisson Sampling:**

$$n_{11} \sim P(\lambda_{11})$$

$$n_{12} \sim P(\lambda_{12})$$

$$n_{21} \sim P(\lambda_{21})$$

$$n_{22} \sim P(\lambda_{22})$$

where λ_{ij} is the expected number of observations falling in cell (i, j).

	Level 1	Level 2	Total
Level 1	λ_{11}	λ_{12}	λ_{1+}
Level 2	λ_{21}	λ_{22}	λ_{2+}
Total	λ_{+1}	λ_{+2}	λ_{++}

Assuming that each of these are independent of the others, the following distributions can be shown

$$R_i \sim P(\lambda_{i+})$$

$$C_j \sim P(\lambda_{+j})$$

$$T \sim P(\lambda_{++})$$

where

$$\lambda_{i+} = \lambda_{i1} + \lambda_{i2}$$

$$\lambda_{+j} = \lambda_{1j} + \lambda_{2j}$$

$$\lambda_{++} = \lambda_{1+} + \lambda_{2+} = \lambda_{+1} + \lambda_{+2}$$

Of more interest are the proportions falling into each cell in the overall table,

$$p_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$$

the row and column marginal proportions

$$\gamma_i = \frac{\lambda_{i+}}{\lambda_{++}}$$
$$= p_{i1} + p_{i2} = p_{i+}$$
$$\rho_j = \frac{\lambda_{+j}}{\lambda_{++}}$$
$$= p_{1j} + p_{2j} = p_{+j}$$

or the conditional proportions following into each row

$$p_{j|i} = P[\text{Column } j | \text{Row } i] = \frac{\lambda_{ij}}{\lambda_{i+}}$$

(given that the observation is from row i, what fraction fall in column j)

or column

$$p_{i|j} = P[\text{Row } i | \text{Column } j] = \frac{\lambda_{ij}}{\lambda_{+j}}$$

(given that the observation is from column j, what fraction fall in row i)

It can be shown that

$$\pi_i = p_{1|i} \qquad \nu_j = p_{1|j}$$

This is based on the fact that with the given Poisson model for the four cells

$$n_{11} | n_{11} + n_{12} = R_1 \sim Bin(R_1, \pi_1)$$

where
$$\pi_1 = \frac{\lambda_{11}}{\lambda_{11} + \lambda_{12}}$$

and

$$n_{11}|n_{11} + n_{21} = C_1 \sim Bin(C_1, \nu_1)$$

where
$$\nu_1 = \frac{\lambda_{11}}{\lambda_{11} + \lambda_{21}}$$

(similarly for the other conditional probabilities.)

This can also be shown by the fact that again with the given Poisson model

$$(n_{11}, n_{12}, n_{21}, n_{22})|\sum n_{ij} = T \sim Multi(T, (p_{11}, p_{12}, p_{21}, p_{22}))$$

where the p_{ij} are as defined as before.

Example: Infant mortality in New York City in 1974

Birthweight	Dead	Alive	Total
$\leq$ 2500 grams	530	4340	4870
> 2500 grams	333	32637	32970
Total	863	36977	37840

This is an example of Poisson sampling as all the births in New York were recorded and classified into the 2×2 table.

So the joint and marginal probabilities are estimated to be

Birthweight	Dead	Alive	Total
$\leq$ 2500 grams	0.0140	0.1147	0.1287
> 2500 grams	0.0088	0.8625	0.8713
Total	0.0228	0.9778	1

$$\frac{530}{37840} = 0.0140 \qquad \frac{32970}{37840} = 0.8713 \qquad \text{etc}$$

The conditional probabilities are estimated to be

$P[\text{Status}|\text{Weight}]$:

Birthweight	Dead	Alive	Total
$\leq$ 2500 grams	0.1088	0.8912	1
$>$ 2500 grams	0.0101	0.9899	1

$P[\text{Weight}|\text{Status}]$:

Birthweight	Dead	Alive
$\leq$ 2500 grams	0.6141	0.1174
$>$ 2500 grams	0.3859	0.8826
Total	1	1

$$\frac{530}{4870} = 0.1088 \qquad \frac{32970}{36977} = 0.8826 \qquad \text{etc}$$

- **Hypergeometric Probability Distribution:**

  In this case, you can consider the situation has sampling from a population of size T, where there are C_1 successes and C_2 failures and you will sample R_1 items from this population, observing the number of successes n_{11}.

  Under the usual null hypothesis assumptions (to be discussed later)

  $$n_{11} \sim Hyper(R_1, C_1, T)$$

  Otherwise, n_{11} has what as known as a non-central hypergeometric distribution

Estimable Parameters and Testable Hypotheses

Sampling Scheme	π_1	π_2	$\pi_1 - \pi_2$	ω_1	ω_2	ω_1/ω_2	ρ	γ
Poisson	✔	✔	✔	✔	✔	✔	✔	✔
Multinomial	✔	✔	✔	✔	✔	✔	✔	✔
Product Binomial								
Prospective	✔	✔	✔	✔	✔	✔	✗	*
Retrospective	✗	✗	✗	✗	✗	✗	*	✗
Randomize Experiment	✔	✔	✔	✔	✔	✔		

* only possible if $\omega_1 = \omega_2$

Lets consider the case of estimating γ in the prospective sampling case when $\omega_1 = \omega_2$. In this case, the two binomial distributions must have the form

	Level 1	Level 2	Total
Level 1	π	$1 - \pi$	1
Level 2	π	$1 - \pi$	1

for some value π.

In this case $P[\text{Col } 1|\text{Row } 1] = P[\text{Col } 1|\text{Row } 2]$, so rows and columns will be independent yielding a joint probability table of

	Level 1	Level 2	Total
Level 1	$r\pi$	$r(1 - \pi)$	r
Level 2	$(1 - r)\pi$	$(1 - r)1 - \pi$	$1 - r$
Total	π	$1 - \pi$	1

where r depends on the given and fixed sample sizes for each row.

This implies that
$$P[\text{Col } 1] = \pi = \gamma$$
for any r.

Then γ can be estimated since

$$C_1 = n_{11} + n_{21} \sim Bin(T, \gamma)$$

Actually the assumption that $\omega_1 = \omega_2$ can be extended to the case that $\omega_2 = \phi\omega_1$ where ϕ is the known (and fixed) odds ratio. Usually ϕ is not known so the case won't be discussed here.

So far, our testing examples we've been looking at null hypotheses of the equivalent forms

$$H_0 : \pi_1 = \pi_2 \qquad \text{and} \qquad H_0 : \omega_1 = \omega_2$$

These are often called hypotheses of homogeneity as they are focused on whether the distribution of the binary response is homogeneous – the same – across populations.

Now consider the situations of Poisson and Multinomial sampling. The above hypotheses are equivalent to the statement

$$H_0 : P[\text{Col 1}|\text{Row 1}] = P[\text{Col 1}|\text{Row 2}]$$

i.e. rows and columns are independent. So under these sampling schemes, tests can either be thought of as tests of independence or tests of homogeneity.

Another way of thinking of the null hypothesis of independence is

$$H_0 : p_{ij} = \rho_i \gamma_j$$

In the product binomial and randomized experiments cases, considering these as tests of homogeneity only make sense as one of the two factors is not random.

For example in the aspirin trial, the number of people of each treatment is fixed. The only randomness is in whether a subject has a stroke or not.

Pearson Chi-Squared Test for 2 × 2 tables

Another way of thinking of testing for independence or homogeneity, is to compare the observed counts in the table with what would be expected under the null hypothesis.

Lets suppose that the sampling mechanism is prospective product binomial sampling. Then as we have seen before, the estimate of the common success probability is

$$\hat{\pi}_c = \frac{S_1 + S_2}{n_1 + n_2}$$

Fitting this into the notation for the 2 × 2 table, this is equivalent to

$$\hat{\pi}_c = \frac{n_{11} + n_{21}}{R_1 + R_2} = \frac{C_1}{T}$$

This implies we should estimate the number of success in the two rows by

$$E_{i1} = R_i \times \hat{\pi}_c = R_i \frac{C_1}{T}$$

Similarly all cells in the table the expected number of counts is

	Level 1	Level 2	Total
Level 1	$\frac{R_1 C_1}{T}$	$\frac{R_1 C_2}{T}$	R_1
Level 2	$\frac{R_2 C_1}{T}$	$\frac{R_2 C_2}{T}$	R_2
Total	C_1	C_2	T

Notice that this table has the same row and column sums as the table of the observed counts.

Now lets think about what to do for Multinomial sampling case under independence. In this case $p_{ij} = \rho_i \gamma_j$, so it seems reasonable to estimate the expected counts by

$$E_{ij} = T \hat{\rho}_i \hat{\gamma}_j$$

Under independence, the marginal probabilities can be estimated by

$$\hat{\rho}_i = \frac{R_i}{T} \qquad \hat{\gamma}_j = \frac{C_j}{T}$$

Using this gives

$$E_{ij} = T \frac{R_i}{T} \frac{C_j}{T} = \frac{R_i C_j}{T}$$

the same expected counts as before.

In fact, for all of the discussed sampling schemes, these are reasonable estimates for the expected counts. (They are all maximum likelihood estimates).

The form of Pearson's Chi-Square test is

$$X^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

This is compared to a χ_1^2 distribution. For this test, always calculate a 2-sided p-value. This statistic does not take account of the direction of the deviation from the null hypothesis.

Why 1 degree of freedom? As noted earlier the table of expected counts must have the row and column sums as the table of observed counts. So once you've specified the marginal totals and one cell count, all other cells are then determined. An equivalent statement is that once $O_{ij} - E_{ij}$ is specified for one cell, is it known for the whole 2×2 table.

So you can think about this in a similar fashion as the degrees of freedom for estimating σ^2 when calculating a t-test statistic.

Lets look to see if birth weight and 1 year mortality are independent.

Observed and expected counts:

	Observed Counts			Expected Counts		
	Dead	Alive	Total	Dead	Alive	Total
≤ 2500g	530	4340	4870	111.0679	4758.932	4870
> 2500g	333	32637	37840	751.9321	32218.068	32970
Total	863	36977	37840	863	36977	37840

$O - E$:

	Dead	Alive
≤ 2500g	418.9321	-418.9321
> 2500g	-418.9321	418.9321

$(O - E)^2/E$:

	Dead	Alive
$\leq$ 2500g	1580.2	36.879
$>$ 2500g	233.4	5.447

$$X^2 = 1855.882 \qquad p\text{-value} < 2.2e - 16$$

A similar answer to last time, there is an association between birth weight and 1 year survival.

This test is implemented in **R** with the `chisq.test()` function. To implement the example with this function

```
> nydeath <- matrix(c(530, 4340, 333, 32637),ncol=2, byrow=T)

> chisq.test(nydeath, correct=F)

        Pearson's Chi-squared test

data:  nydeath X-squared = 1855.882, df = 1, p-value < 2.2e-16
```

So we now have three tests for investigating tests on proportions

- z-test based on $\hat{\pi}_1 - \hat{\pi}_2$

- z-test based on log odds ratio

- Pearson Chi-square test

As mentioned before the two z-tests will give similar, but not exactly the same answers. However it can be shown that for the test based on $\hat{\pi}_1 - \hat{\pi}_2$

$$X^2 = z^2$$

Since the square of a $N(0,1)$ is distributed χ_1^2, the p-value must be the same for both tests (assuming 2-sided z-test).

Since these two tests are the same, we should use the same sample size considerations to determine whether the asymptotic χ_1^2 distribution approximation is reasonable. The usual assumption can be thought of in this setting as

$$\min E_{ij} \geq 5$$

If this condition is not met, there is an approximation due to Yates that may be used. The test statistics is modified to

$$X_y^2 = \sum_{\text{all cells}} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

This is the default case with the function `chisq.test` (which is way the earlier example had the `correct=F` option set).

For large examples, the correction won't make much of a difference as can be seen in

```
> chisq.test(nydeath, correct=F)

    Pearson's Chi-squared test

data:  nydeath X-squared = 1855.882, df = 1, p-value < 2.2e-16

> chisq.test(nydeath, correct=T)

    Pearson's Chi-squared test with Yates' continuity correction

data:  nydeath X-squared = 1851.454, df = 1, p-value < 2.2e-16
```

Though for smaller examples, it may. Consider the urethritis in HIV-I example

```
> chisq.test(SSS, correct=F)

    Pearson's Chi-squared test

data:   SSS X-squared = 8.4156, df = 1, p-value = 0.00372

Warning message: Chi-squared approximation may be incorrect in:
chisq.test(SSS, correct = F)

> chisq.test(SSS, correct=T)

    Pearson's Chi-squared test with Yates' continuity correction

data:   SSS X-squared = 5.0909, df = 1, p-value = 0.02405

Warning message: Chi-squared approximation may be incorrect in:
chisq.test(SSS, correct = T)
```

To see what happens with a situation of the boundary of the rule of thumb, consider the dataset analyzed by Fisher (1950)

Twin Type	Convicted	Not convicted
Dizygotic	2	15
Monozygotic	10	3

The expected counts are

Twin Type	Convicted	Not convicted
Dizygotic	6.8	10.2
Monozygotic	5.2	7.8

```
>  chisq.test(Convictions, correct=F)
```

Pearson's Chi-squared test

data: Convictions X-squared = 13.0317, df = 1, p-value = 0.0003063

```
> chisq.test(Convictions, correct=T)
```

Pearson's Chi-squared test with Yates' continuity correction

data: Convictions X-squared = 10.4581, df = 1, p-value = 0.001221

while the p-value changes by a factor of about 4, the effective answer is the same.

There is also a z-test version of Yates correction, where

$$z_y = \frac{\left|\hat{\pi}_1 - \hat{\pi}_2\right| - \frac{1}{2T}}{\sqrt{\frac{\hat{\pi}_c(1-\hat{\pi}_c)}{T}}}$$

The general consensus today is not to use Yates correction, as it is too conservative and leads to tests with lower power.

If you are worried about the asymptotics going into calculation of the p-value, you are better using one of the procedures to be discussed next class.

Extension of Chi-square tests:

While the Chi-square test was displayed for 2×2 tables, it can easily be extended to $r \times c$ tables. This will be discussed later when we get to the log-linear model.