UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor**: **Ivo Dinov,**
  Asst. Prof. In Statistics and Neurology

- **Teaching Assistants:** Tom Daula and Ming Zheng
  UCLA Statistics

University of California, Los Angeles, Winter 2003

http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 11: Tables of Counts

We discussed means and mean differences in Ch. 10 and developed a statistical toolbox for analyzing quantitative variables.

Now we want to develop a similar approach for analyzing qualitative variables.

Table-of-measurements → tables-of-counts;

Means → proportions

T/F-tests for inference on qualitative variables →

Chi-square (χ^2) tests for categorical data.

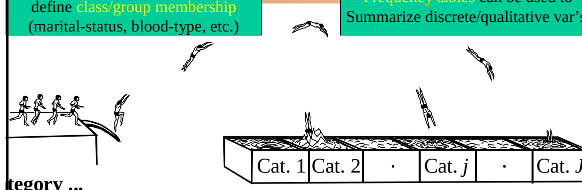
Chapter 11: Tables of Counts

- One-dimensional tables
  and goodness of fit
- Two-way tables of counts
  Chi-square test of homogeneity
  Chi-square test of independence
  2 by 2 tables
- The perils of collapsing tables

1-dimensional tables – classify *n*-individuals in *J*-categories

Qualitative (factors), class variables define class/group membership (marital-status, blood-type, etc.)

Frequency tables can be used to Summarize discrete/qualitative var's.



Probability	p_1	p_2	· · ·	p_j	· · ·	p_J
Observed count	O_1	O_2	· · ·	O_j	· · ·	O_J
Expected count	E_1	E_2	· · ·	E_j	· · ·	E_J

$$E_j = n\,p_j$$

1-dimensional tables cont.

expected cell count = total × specified cell probability

The T and F statistics are used for inference about quantitative variables. χ^2 statistics is used for analysis of categorical data.

- When H_0 gives the probabilities of landing in each cell completely (no parameters to be estimated) , $P(cell_1)=p_1$, $P(cell_2)=p_2$, …, $P(cell_J)=p_J$, and $\Sigma p_k=1$.
- Thus, having J-1 probabilities fixed determines the last probability.

df = number of categories - 1

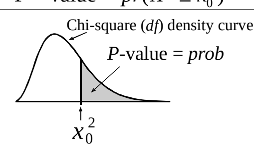
Chi-Square Test – goodness of fit test

- The Chi-square test statistic (χ^2) has observed value

$$x_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed - expected})^2}{\text{expected}}$$

- The *P*-value for the test is

$$P-\text{value} = pr(X^2 \geq x_0^2) \quad \text{where } X^2 \sim \text{Chi-square}(df)$$



To test a null-hypothesis, H_0, we compare the observed counts in the table to the expected (theoretical) counts. For this reason this test is called a goodness-of-fit test – observed/expected count fit.

Example of 1D table – Three blood types

TABLE 11.1.1 Proportions of Three Blood Types

	A	AB	B	Total
No. Observed	39	70	42	151
Proportion Observed	0.258	0.464	0.278	1.000

Example of 1D table – rolling a single die

TABLE 11.1.2 210 Rolls of a Die

Outcome	1	2	3	4	5	6	Total
Count	26	40	37	26	43	38	210
Proportion	0.124	0.190	0.176	0.124	0.205	0.181	1.000

Why aren't these probabilities all equal?!?
Are they supposed to?
What are the expected probabilities (1/6)?
χ^2 statistics is x_0=7.54, df=5, P-value=0.18

Exit poll – sampling voters as they leave polling booths. Exit polls of 10,000 voters.

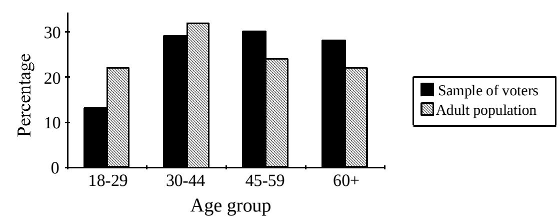
(a) Table of exit-poll sample and population Age distributions

		Age group				
		18-29	30-44	45-59	60+	Total
Sample :	*(Percentages)*	13	29	30	28	100
Population :	*(Percentages)*	22	32	24	22	100

Are there differences between the population age groups and the exit-poll sample age groups?
Younger voter underrepresented voters.
Real differences or just due to sampling variation?

Exit poll – Bar-plot of population/sample groups

(b) Plot of exit-poll sample and population Age distributions



H_0: True proportions in the 4 age groups in the voter sample and the whole population are the same!

Exit poll – Bar-plot of population/sample groups

(c) Table of observed and expected counts

	Age group				
	18-29	30-44	45-59	60+	Total
Observed count	1300	2900	3000	2800	10,000
Expected count	2200	3200	2400	2200	10,000

(Note: Counts approximate due to the rounding of percentages in the report.)

Figure 11.1.1 Comparing the age distributions for voters and the population.

H_0: $p_{18\text{-}29} = 0.22$; $p_{30\text{-}34} = 0.32$; $p_{45\text{-}59} = 0.32$; $p_{60+} = 0.32$;

Exit poll – Bar-plot of population/sample groups

$$x_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = 709.94$$

$$\text{df} = \text{number of groups} - 1 = 4 - 1 = 3$$

P-value = 0.000, very small, indicating extremely strong evidence against the null-hypothesis. The 95% CI for each age groups are:
[12.3 : 13.7]; [28.1 : 30.0]; [29.1 : 30.9]; [27.1 : 28.9]

The Chi-square distribution

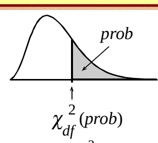


Figure 11.1.3 The $\chi^2_{df}(prob)$ notation

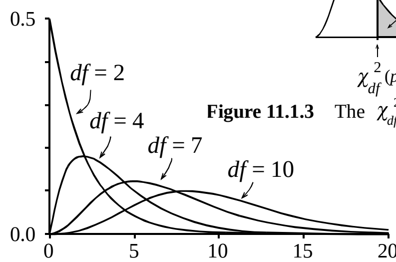


Figure 11.1.2 Chi-square(*df*) p.d.f. curves.

Lotto after 399 numbers have been drawn – Do some numbers appear more frequently in LOTTO?

TABLE 11.1.3 Frequency of Winning Numbers in LOTTO

1.	(7)	**2.**	(10)	**3.**	(8)	**4.**	(9)	**5.**	(13)	**6.**	(8)	**7.**	(12)	**8.**	(16)	**9.**	(11)	**10.**	(6)
11.	(13)	**12.**	(10)	**13.**	(9)	**14.**	(11)	**15.**	(11)	**16.**	(6)	**17.**	(11)	**18.**	(13)	**19.**	(6)	**20.**	(13)
21.	(7)	**22.**	(9)	**23.**	(8)	**24.**	(12)	**25.**	(6)	**26.**	(4)	**27.**	(10)	**28.**	(8)	**29.**	(14)	**30.**	(12)
31.	(11)	**32.**	(12)	**33.**	(9)	**34.**	(11)	**35.**	(6)	**36.**	(8)	**37.**	(14)	**38.**	(10)	**39.**	(15)	**40.**	(10)

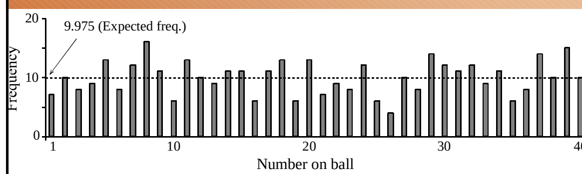


Figure 11.1.4 Frequency of LOTTO winning numbers

Lotto after 399 numbers have been drawn – Do some numbers appear more frequently in LOTTO?

Number-range: [1:40]
Number of balls selected at each draw: 7
Number of samples: 57
Total number of balls selected: 57*7=399,
Expected value of each number: 399/40 = 9.975
Observed χ^2 statistics is x_0=30.97
df=40-1=39
P-value = 0.817
Conclusion: No evidence for departure from the null hypothesis.

Review

1. The test statistic for the Chi-square test compares observed and expected frequencies. In what sense are the *expected* frequencies expected? (Expected frequencies are the frequencies expected in H_0 were true.)
2. What shape does the Chi-square distribution generally have? What happens to its shape as the degrees of freedom increase? (Skewed unimodal, becomes symmetric and Normal approximates it well for large df.)
3. What values of the Chi-square test statistic (large or small) provide evidence against the null hypothesis? Why? (Large values, since P-value is small. See density curve.)

Review

4. For one-dimensional tables, how do you compute the degrees of freedom *df* ? (df=number of cells/groups-1.)
5. Do the expected counts have to be whole numbers? (No, expected counts = number of samples x cell-probability.)

Two-way tables

Suppose we have two (or more) qualitative variables, that we use to classify individuals/units/subjects into groups/classes.

Example, 400 patients with malignant melanoma (type of skin cancer) are cross-classified by TYPE (malignant-cell-type) and SITE (focal-location).

4x3 table (4-rows, types and 3 columns, sites).

Questions: What's the most common type of cancer? What locations are mostly effected?

Melanoma by Site and Type, e.g., 4.6.2

TABLE 11.2.1 Four hundred Melanoma Patients by Type and Site

Type	SITE: Head and neck	Trunk	Extremities	Row totals
Hutchinson's Melanomic Freckle	22	2	10	34
Superficial Spreading Melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Column Totals	68	106	226	400

Source: Roberts et al. [1981]

TABLE 11.2.1 Four hundred Melanoma Patients by Type and Site

Type	SITE: Head and neck	Trunk	Extremities	Row totals
Hutchinson's Melanomic Freckle	22	2	10	34
Superficial Spreading Melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Column Totals	68	106	226	400

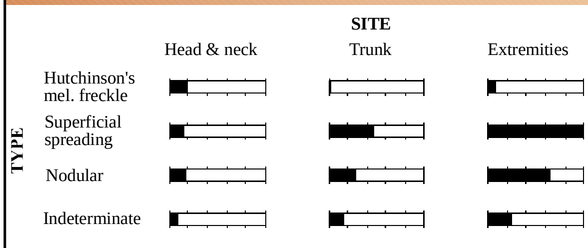
Proportions of all 400 patients

entry = count/400

TABLE 11.2.2 Proportions of Melanoma Patients by Type and Site (whole-table proportions)

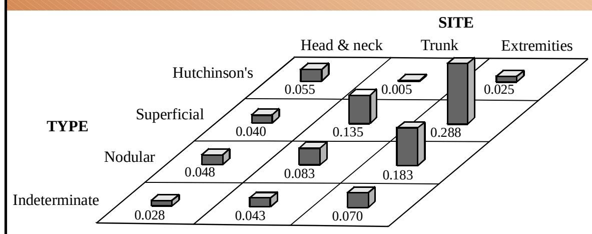
Type	SITE: Head and neck	Trunk	Extremities	Row totals
Hutchinson's Melanomic Freckle	0.055	0.005	0.025	0.085
Superficial Spreading Melanoma	0.040	0.135	0.288	0.463
Nodular	0.048	0.083	0.183	0.313
Indeterminate	0.028	0.043	0.070	0.140
Column Totals	0.170	0.265	0.565	1.000

Remember plot/graph/look at your data first!



(a) Thermometer plot

Remember plot/graph/look at your data first!



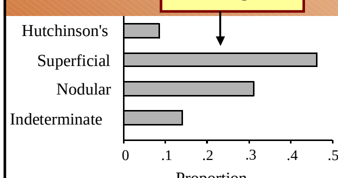
(b) 3-dimensional plot

Next, summarize row/columns – distributions by Type/Site

TABLE 11.2.2 Proportions of Melanoma Patients by Type and Site (whole-table proportions)

Type	SITE: Head and neck	Trunk	Extremities	Row totals
Hutchinson's Melanomic Freckle	0.055	0.005	0.025	0.085
Superficial Spreading Melanoma	0.040	0.135	0.288	0.463
Nodular	0.048	0.083	0.183	0.313
Indeterminate	0.028	0.043	0.070	0.140
Column Totals	0.170	0.265	0.565	1.000

From the row totals we get ….



From the col totals we get ….

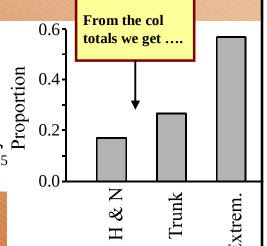


TABLE 11.2.2 Proportions of Melanoma Patients by Type and Site (whole-table proportions)

Type	SITE: Head and neck	Trunk	Extremities	Row totals
Hutchinson's Melanomic Freckle	0.055	0.005	0.025	0.085
Superficial Spreading Melanoma	0.040	0.135	0.288	0.463
Nodular	0.048	0.083	0.183	0.313
Indeterminate	0.028	0.043	0.070	0.140
Column Totals	0.170	0.265	0.565	1.000

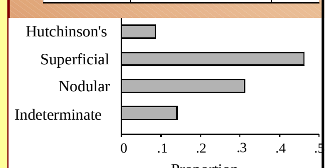
Next, analyze the data-

1. Goodness-of-fit χ^2-test, testing equality of the **TYPE proportions**.

x^2_0=141.42, df=3, P-value $<10^{-6}$

Cancer types are not equally likely!

Can we address questions about the **Site and Type of cancer simultaneously**?



Example - Blood types

TABLE 11.2.3 Regional Data for the ABO System

REGION	PHENOTYPE A	B	O	AB	Total
Nithsdale	98	35	115	5	253
Cree	38	9	79	6	132
Rhinns	36	9	47	7	99
Total	172	53	241	18	484

TABLE 11.2.4 ABO Distributions from Three Areas (Row proportions)[a]

REGION	PHENOTYPE A	B	O	AB	Total
Nithsdale	0.39	0.14	0.45	0.02	1.00
Cree	0.29	0.07	0.60	0.05	1.00
Rhinns	0.36	0.09	0.47	0.07	1.00

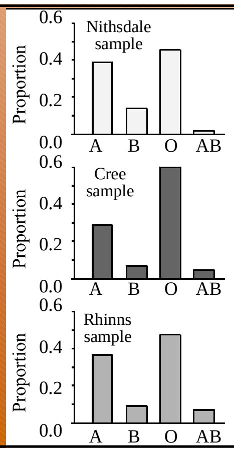
Example - Blood types

Blood contains genetic info that can help determine if populations in some Geo-regions have different racial origins from those in other regions.

This is blood donor data from SW Scotland, Mitchell, 1976. Data (obs. study) is classified using the ABO type system in a 3x4 table (region/phenotype).

Q: Are there regional differences in the phenotype structure?

Assume: random sample from real population, w.r.t. the ABO blood type.



Row distributions

TABLE 11.2.4 ABO Distributions from Three Areas (Row proportions)

REGION	PHENOTYPE A	B	O	AB	Total
Nithsdale	0.39	0.14	0.45	0.02	1.00
Cree	0.29	0.07	0.60	0.05	1.00
Rhinns	0.36	0.09	0.47	0.07	1.00

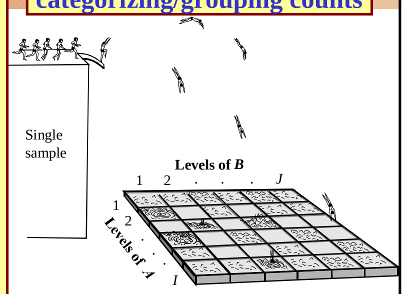
Blood-type O – most common

Two-way tables – 1 sample categorizing/grouping counts

Since there appears to be visual differences in proportions of A-type, between sites.

We can compare proportions of people in Nithsdale with type A to the proportions of people in Cree with blood type A.

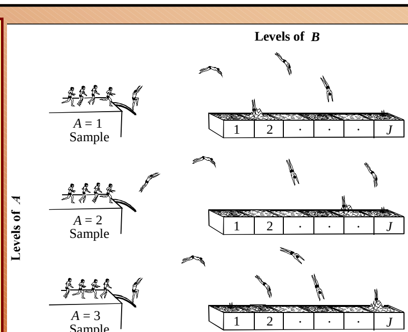
Use difference in proportions, Sec. 8.5 H_0: $p_1 - p_2 = 0$



Situation (1): One sample cross-classified by 2 factors

H_0: No relationship between the level of *B* fall in the level of *A* fall in. Independence?

Two-way tables – many samples diff-levels of a single response factor



Situation (2): *I* samples, single response factor

H_0: the distribution of *B* levels is the same for every level of *A*. Homogeneity?

Notation

TABLE 11.2.5 General Notation for a Two-Way Table

Level of A	Level of B: 1	2	…	j	…	J	Total
1	O_{11}	O_{12}	…	O_{1j}	…	O_{1J}	R_1
2	O_{21}	O_{22}	…	O_{2j}	…	O_{2J}	R_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{iJ}	R_i
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
I	O_{I1}	O_{I2}	…	O_{Ij}	…	O_{IJ}	R_I
Total	C_1	C_2	…	C_j	…	C_J	n

Chi-Square Test
(for either homogeneity or independence)

- The Chi-square test statistic has observed value

$$x_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i,j} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

$$\hat{E}_{ij} = \frac{R_i C_j}{n} = \frac{i\text{ th row total} \times j\text{ th col total}}{\text{grand total}}$$

and $df = (I-1)(J-1)$

- The *P*-value for the test is

$P-\text{value} = pr(X^2 \geq x_0^2)$ where $X^2 \sim \text{Chi-square}(df)$

Chi-square test output – Cancer Type/Site

Chi-Square Test

```
Expected counts are printed below observed counts
                  Head & N    Trunk   Extremit    Total
               1        22        2         10       34
Hutchinson's          5.78     9.01      19.21
               2        16       54        115      185
Superficial          31.45    49.03     104.53
               3        19       33         73      125
Nodular              21.25    33.13      70.62
               4        11       17         28       56
Indeterminate         9.52    14.84      31.64
Total                   68      106        226      400
Chi-Sq = 45.517 +  5.454 +  4.416 +
          7.590 +  0.505 +  1.050 +
          0.238 +  0.000 +  0.080 +
          0.230 +  0.314 +  0.419 = 65.813
DF = 6, P-Value = 0.000
```

Figure 11.2.5 Minitab output for the melanoma data.

Comments

1. What information do the row sums of a 2-way table of counts give you? What about the column sums?
2. How do you calculate whole-table proportions? When does it make sense to calculate them? What information do such proportions give you?
3. What sort of information do the row sums of the whole-table proportions give you? What about the column sums?
4. What are the denominators of the row proportions? What information do they give you? Repeat for column proportions.

Comments

5. Suppose that we are interested in comparing row distributions. In what way(s) can we sample to obtain our data? Express in words the null hypothesis tested by the Chi-square test. Repeat for column distributions.
6. If we do not want to think in terms of row distributions or column distributions but just want to see whether there is any relationship between the row an individual falls into and the column he or she falls into, express in words the null hypothesis tested by the Chi-square test.

Comments

7. Express in words how one calculates the expected counts for cell(*i, j*).
8. Qualitatively, would a large value or a small value of x_0^2 make you think that there was evidence of a relationship between row and column distributions? Why?
9. Qualitatively, would a large *P*-value or a small *P*-value make you think that there was evidence of a relationship between row and column classifications?

Degrees of freedom – since there are n-1 free parameters, for colums and rows, row/comun sums must equal 1 (or n)

Chi-square test for a 2 × 2 table: ***df* = 1**.

In general for *IxJ* table *df=(I-1)*(J-1)*

Dangers of collapsing tables -Simpson's paradox

	Nonsmokers			Smokers	
	Not irradiated	Irradiated		Not irradiated	Irradiated
No cancer	950	9000	No cancer	5000	5
Cancer	50(5%)	1000(10%)	Cancer	5000(50%)	95(95%)

TABLE 11.3.3 The Collapsed Table

	Not irradiated	Irradiated
No cancer	5950	9005
Cancer	5050(46%)	1095(11%)

Collapsing the 3-way table (artificial, on top) to a 2-way table (bottom), w.r.t. smoking factor. Goal to investigate irradiation/cancer relation. It appears as though irradiation decreases the cancer rate …

Chapter 11 Summary

General Ideas about Chi-Square Tests

- The Chi-square test statistic has observed value

$$x_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed - expected})^2}{\text{expected}}$$

- The *P*-value for the test is

$$P-\text{value} = pr(X^2 \geq x_0^2) \quad \text{where } X^2 \sim \text{Chi - square}(df)$$

Chi-square tests cont.

- **Observed** refers to the count observed in the cell (i.e. what the data says).
- **Expected** refers to the count that would be expected if H_0 was true.
- **Large values** of x_0^2 provide evidence against H_0. Such values occur when we get observed counts far from what H_0 would lead us to expect.
- The **degrees of freedom *(df)*** depend on the dimension(s) of the table and the hypothesis being tested.
- The individual terms in the sum (one for each cell) are called the components of Chi-square. When we have a statistically significant test result, inspecting the large components can lead to insight into how the hypothesis is failing to describe the data.

Warning about Chi-square

- Using the Chi-square distribution as the sampling distribution of X^2 when H_0 is true is a large sample approximation.
- Where expected counts are small, *P*-values from the Chi-square distribution begin to become unreliable.
- Our rule is that expected counts should be greater than 1 and 80% of the expected counts should be at least 5.
- If this rule is not satisfied, we can often amalgamate rare categories
  - (i.e. treat two or more similar classes as a single class) in order to increase the expected counts.
- For 2 x 2 tables we use the rule for comparing two proportions.

One-Dimensional Tables

- A single sample of units or individuals being classified into groups by a single factor (with *J* levels).
  - We summarize the data using a (1-way) frequency table and plot it using a bar graph.
  - Chi-squared tests are useful when we have a hypothesis defining the values of the set of probabilities (or population proportions) that the data was sampled from.
  - The degrees of freedom is $df = J\text{-}1$
    - this applies if the set of probabilities is completely specified
    - if the probabilities are hypothesized to come from a distribution with parameter(s) that must estimated from the data then $df = k - 1 - \#\text{estimated parameters}$

One-dimensional tables cont.

- A common hypothesis is that all of the probabilities (respectively population proportions) are identical.
- If the above hypothesis is rejected, we can investigate the nature of the differences by looking at the differences between pairs of proportions.
- When constructing confidence intervals for differences between proportions, use standard errors for single sample and several response categories.

Two-Way Tables

Chi-Square test

- Whether H_0 specifies *equality of row distributions, or equality of column distributions, or independence of row and column classifications,* the Chi-square test uses
  Expected count in cell(i,j):

$$\hat{E}_{ij} = \frac{R_i C_j}{n} = \frac{i\text{th row total} \times j\text{th col total}}{\text{grand total}}$$

and

$$df = (I-1)(J-1)$$

Warning

- Chi-square tests, as described in this book, are only appropriate when the data is collected as a single random sample or when rows (or columns) come from independent random samples.
- Social scientists have often used it on two-way tables constructed using data from complex surveys which employ devices such as cluster sampling.
- The Chi-square test is not appropriate under such circumstances.

Two-way tables cont.

Two Types of Table

- We distinguished between
  - Situation 1, Single sample cross-classified by two factors
  - and Situation (2), separate samples, each classified according to one response factor (see Fig.11.2.7).

Row distributions

- Row distributions tell us about the chances that an individual who belongs to a given row will fall into each of the column classes.
- They are estimated by the row proportions of the table (using row totals as denominators).
- They are not meaningful if columns are separate samples.
- When constructing confidence intervals for differences between proportions, proportions from different rows are statistically independent.

Column Distributions

- Column distributions tell us about the chances that an individual who belongs to a given column will fall into each of the row classes.
- They are estimated by the column proportions of the table (using column totals as denominators).
- They are not meaningful if rows are separate samples.
- When constructing confidence intervals for differences between proportions, proportions from different columns are statistically independent.

Whole-table Proportions

- Whole-table proportions are formed using the grand total of the table as the denominator.
- They tell us about the chances of an individual experiencing a given combination of the 2 factors.
- They are only meaningful when we have a single sample cross-classified by two factors.
  - They are not meaningful if rows are separate samples or if columns are separate samples.
- When constructing confidence intervals for differences between proportions, use standard errors for single sample, several response categories.