Chapter 16: Analysis of contingency tables

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- Contingency tables from grouping categorical data
- Modelling contingency tables using `family=binomial`
- Modelling contingency tables using `family=poisson`
- The equivalence of the binomial and Poisson models
- A new interpretation of odds
- Odds ratios (optional section)
- Chi-square test of association (optional section)

Section 16.1
Introduction

Categorical data

Count data often arise from the observation of categorical data.

Data are said to be "categorical" if the measurements made on each subject are ALL factor variables.

The levels of the factor variables are the "categories", that is, they are the distinct values that the factor variable can take.

The counts are then the number of times (i.e., frequencies) each combination of factor levels occurs.

The counts can be arranged in the form of a contingency table.

Categorical data example...

Vaccine study

Suppose that two vaccine treatments (Trmt A and Trmt B) are to be compared for local tenderness around the injection site. Each subject receives one of the two vaccines, and the degree of local tenderness is classified into one of four levels.

- no tenderness
- mild tenderness
- moderate tenderness
- severe tenderness

The number of each treatment-tenderness combination can then be counted and presented in the form of a contingency table:

	none	mild	moderate	severe
Trmt A	21	16	11	2
Trmt B	1	22	19	9

Categorical data example

Hair and eye colour study

A genetic study wanted to determine if eye colour is associated with hair colour.

A large collection of randomly chosen online portrait photographs were examined and hair and eye colour was classified as brown or not brown (other). The resulting contingency table was

	Eyes brown	Eyes other
Hair brown	284	613
Hair other	577	2002

Categorical data. . .

Attendance/Pass

It is of interest to examine whether attendance of STATS 20x lectures had an association with success (pass or fail) in the course. Recall, we have data for the class of 146 students.

The raw format for recording categorical data is the usual rectangular format with rows being the observations on each subject, and columns being the measured factor variables. The first 8 lines of the raw data look like this:

```
> AP.df = read.table("Data/AttendPass.txt",header=T)
> head(AP.df, 8)
  Subject Pass     Attend
1       1 pass     attend
2       2 pass     attend
3       3 pass     attend
4       4 pass     attend
5       5 pass     attend
6       6 fail not.attend
7       7 pass not.attend
8       8 fail     attend
```

Categorical data. . .

Attendance/Pass. . .

The contingency table of counts can be obtained using the `table` function.

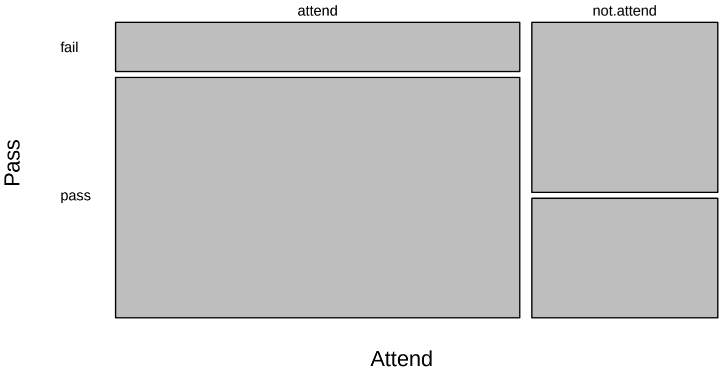
```
> AP.tbl=with(AP.df,table(Attend,Pass))
> AP.tbl
            Pass
Attend       fail pass
  attend       17   83
  not.attend   27   19
```

Categorical data. . .

Attendance/Pass – plotting the counts

It is easy to get some useful plots of the contingency table. Since AP.tbl is a table, the plot function produces a mosaic plot:

```
> plot(AP.tbl,main="",las=1)
```



Note that the size of the rectangles is proportional to the count value.

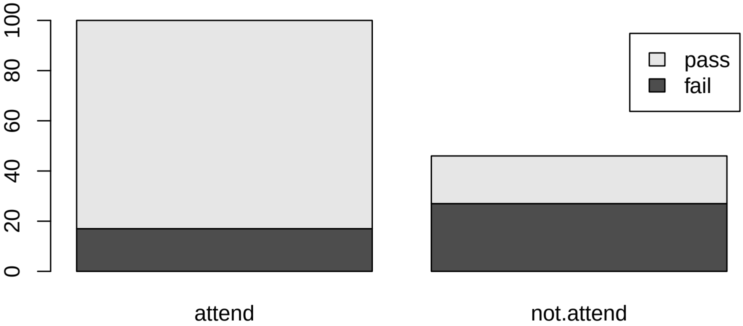
What does this plot tell us?

Categorical data. . .

Attendance/Pass – plotting the counts . . .

The barplot function provides a useful bar plot:

```
> barplot(t(AP.tbl),legend=T)
```

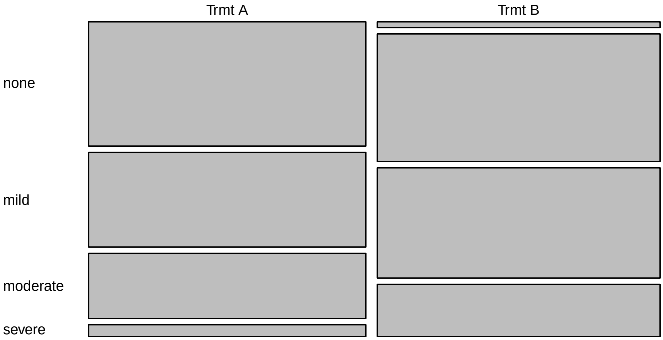


In the above code we used the transpose function t to flip the table rows and columns so that each bar would be for a level of Attend. If we hadn't done this the bars would have been for the levels of Pass, which makes interpretation harder.

Categorical data. . .

Vaccine tenderness – plotting the counts

```
> vaccines=matrix(c(21,16,11,2,1,22,19,9),nrow=2,byrow=T)
> rownames(vaccines)=c("Trmt A","Trmt B")
> colnames(vaccines)=c("none","mild","moderate","severe")
> vax.tbl=as.table(vaccines)
> vax.tbl
       none mild moderate severe
Trmt A   21   16       11      2
Trmt B    1   22       19      9
> plot(vax.tbl,main="",las=1)
```

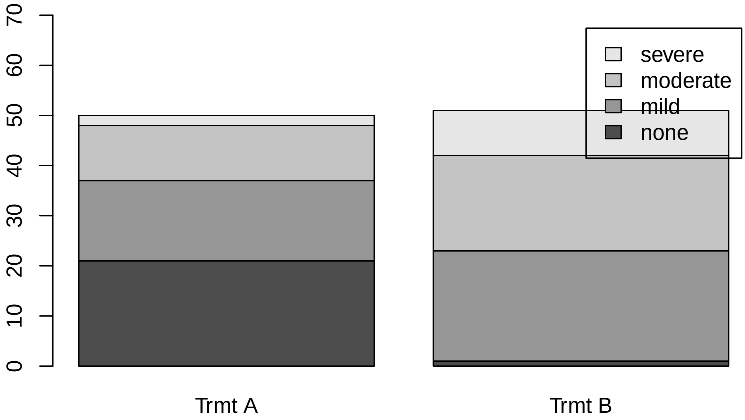


Categorical data. . .

Vaccine tenderness – plotting the counts . . .

The barplot function is not very good at placing the legend!

```
> barplot(t(vax.tbl),ylim=c(0,70),legend=T)
```



Remark 1

In each of the above examples the frequencies were presented in the form a two-way (i.e., by row and by column) contingency table.

The contingency table was 2-by-4 for the vaccine study, and 2-by-2 in the hair/eye colour and attendance/pass example. In general, the number of rows and columns in a two-way contingency table is given by the number of levels of the corresponding factors.

More generally, if s factor variables are recorded on each subject then the resulting table will be a s-way contingency table, with dimensions given by the number of levels of each of the s factors.

Remark 2

With categorical data, it may or may not be possible to identify some of the factor variables as explanatory variables and some as response variables.

It totally depends on the particular situation:

- The vaccine study was conducted to see if tenderness depends on which vaccine treatment was given, so it is clear that treatment is the explanatory variable, and the measured tenderness is the response.
- In the hair/eye colour example neither is clearly an explanatory or response.
- In the attendance/pass example it is natural to consider attendance as an explanatory variable for pass, and indeed we have already done that in this course.

Remark 3

With categorical data, it may or may not be possible to say whether the counts are from a fixed number of trials (e.g., binomial data), or are more like Poisson data.

- In the vaccine example it is most likely that there was a fixed number of treatments applied, so if we created a dichotomous response (such as, "no tenderness" or "some tenderness") then we could model it as binomial.
- In the hair/eye colour example there is no clear notion of a fixed number of trials for any hair or eye colour.
- In the attendance/pass example:
  - One could say that the number of attenders and non-attenders was fixed (at 100 and 46, respectively) prior to the exam.
  - Alternatively, one could argue that the the number of attenders and non-attenders was not fixed since it depended on the number of enrolments.

Remarks

The last two slides above note some interesting properties of categorical data. As we shall see, these considerations don't matter to our analysis – but they may determine our interpretation.

The underlying research question is to establish whether or not there is an association between the factors.

If there is an association then we would also like to be able to quantify it.

NOTE: This Chapter concludes with a recap of the methodology seen in STATS 10x for testing for association in contingency tables – the chi-squared test for association. Your lecturer will advise whether it is examinable.

Section 16.2
The binomial approach to contingency table analysis

Attendance/Pass

Two STATS 20x students, Kim and Des, have been assigned to the task of determining whether there is an association between attendance and exam success in STATS 20X.

Kim has decided to use the binomial approach to analyse the data – this makes sense, since we can regard attendance as an explanatory variable, and pass/fail as a Bernoulli outcome.

We'll help Kim out by creating a dataframe in the format needed for a binomial GLM.

```
> Freqs.df = data.frame(Attend=c("not.attend","attend"),Fail=c(27,17),Pass=c(19,83)
> Freqs.df = transform(Freqs.df,Attend=factor(Attend))
> Freqs.df
      Attend Fail Pass
1 not.attend   27   19
2     attend   17   83
```

If an association is detected then Kim also needs to quantity the strength of the assocation with a suitable confidence interval.

Attendance/Pass. . .

Binomial analysis

Kim decides to change the reference level of `Attend` to `not.attend` so that any effect can be expressed as the effect of attending.

Fitting the binomial GLM is easy:

```
> AP.binom=glm(cbind(Pass,Fail)~Attend,data=Freqs.df,family=binomial)
> summary(AP.binom)
```

```
Call:
glm(formula = cbind(Pass, Fail) ~ Attend, family = binomial,

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.3514     0.2994  -1.173    0.241
Attendattend   1.9370     0.4007   4.834 1.34e-06 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  2.5162e+01  on 1  degrees of freedom
Residual deviance: -4.2188e-15  on 0  degrees of freedom
```

Attendance/Pass...

Binomial analysis...

Note that the residual deviance is zero, on zero degrees of freedom – this is nothing to worry about.[1]

The effect of `Attend` is highly significant, so let's calculate a CI:

```
> exp(confint(AP.binom))[2,]
Waiting for profiling to be done...
    2.5 %    97.5 %
 3.214049 15.552487
```

So, Kim concludes that the odds of an attender passing STATS 20x are between 3.2 and 15.6 times the odds of a non-attender passing. Yikes!

[1] The data consist of two observations, and the model fits two parameters, so there are no degrees of freedom left.

Section 16.3
The Poisson approach to contingency table analysis

Attendance/Pass. . .

Poisson analysis

The other student, Des, feels that the frequencies are best modeled as Poisson counts because the number of students was not fixed at the start of the semester, and depended on the whims of enrolment.

Des needs the data in the following format:

```
> library(dplyr)
> AP.df = read.table("Data/AttendPass.txt",header=T)
> AP.df = transform(AP.df, Pass=factor(Pass), Attend=factor(Attend))
> Freqs2.df = AP.df %>% group_by(Attend,Pass) %>% summarize(freq=n()) %>%
+                    data.frame()
> Freqs2.df
      Attend Pass freq
1     attend fail   17
2     attend pass   83
3 not.attend fail   27
4 not.attend pass   19
```

Attendance/Pass. . .

Poisson analysis. . .

Des fits the interaction model to check for an association between attendance and passing.[2] He also relevels `Attend`.

```
> Freqs2.df$Attend=relevel(Freqs2.df$Attend, ref="not.attend")
> AP.pois=glm(freq~Attend*Pass,family=poisson,data=Freqs2.df)
> summary(AP.pois)
```

```
Call:
glm(formula = freq ~ Attend * Pass, family = poisson, data = Freqs2.df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             3.2958     0.1925  17.126  < 2e-16 ***
Attendattend           -0.4626     0.3096  -1.494    0.135
Passpass               -0.3514     0.2994  -1.173    0.241
Attendattend:Passpass   1.9370     0.4007   4.834 1.34e-06 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  6.9305e+01  on 3  degrees of freedom
Residual deviance: -3.9968e-15  on 0  degrees of freedom
```

[2] That is, to see if the expected numbers in the `pass` group relative to the `fail` group depend on attendance.

Attendance/Pass...

Poisson analysis...

As with the binomial model,the residual deviance is zero, on zero degrees of freedom. Again, this is nothing to worry about.[3]

Note that the interaction effect is highly significant. In fact, **the `Attend:Pass` interaction effect from Des's Poisson analysis is exactly the same as the `Attend` effect from Kim's binomial analysis**.

The CI for $\exp(\beta_3)$ from the Poisson model is identical to Kim's CI

```
> exp(confint(AP.pois))[4,]
Waiting for profiling to be done...
    2.5 %    97.5 %
 3.214049 15.552487
```

The next section shows why these two different models produce the same estimate.

[3] The data consist of four observed counts, and the model fits four parameters, so there are no degrees of freedom left.

Section 16.4
Equivalence of the binomial and Poisson approaches

Equivalence of binomial and Poisson approaches

Consider the 2-by-2 contingency table:

$$\begin{array}{cc} n_{11} & n_{12} \\ n_{21} & n_{22} \end{array}$$

We'll assume that row 1 and column 1 correspond to the reference levels for the row and column factor variables, respectively.

Then, from Chapters 13 and 14, the Poisson interaction model[4] for these data is

$$\begin{aligned} \log \mathrm{E}[n_{11}] &= \beta_0 \\ \log \mathrm{E}[n_{21}] &= \beta_0 + \beta_1 \\ \log \mathrm{E}[n_{12}] &= \beta_0 + \beta_2 \\ \log \mathrm{E}[n_{22}] &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \end{aligned}$$

[4] Here we assume the model `glm(n ∼ RowVar * ColVar, family=poisson)` where `RowVar` and `ColVar` are the row and column factor variables, respectively.

Equivalence of binomial and Poisson approaches...

An odds interpretation...

Equivalently,

$$\begin{aligned}
\mathrm{E}[n_{11}] &= \exp(\beta_0) \\
\mathrm{E}[n_{21}] &= \exp(\beta_0 + \beta_1) \\
\mathrm{E}[n_{12}] &= \exp(\beta_0 + \beta_2) = \mathrm{E}[n_{11}] \times \exp(\beta_2) \\
\mathrm{E}[n_{22}] &= \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3) = \mathrm{E}[n_{21}] \times \exp(\beta_2 + \beta_3)
\end{aligned}$$

So, we can say that for every one occurrence expected in the $[1,1]$ cell we expect $\exp(\beta_2)$ occurrences in the $[1,2]$ cell.

Another way of interpreting this is that within row 1, an occurrence is $\exp(\beta_2)$ times as likely to be in column 2 than column 1.

What we've just said is that, within row 1, the odds of being in column 2 (rather than column 1) is $\exp(\beta_2)$.[5]

[5] Recall this example from Chapter 15: If the odds of A are 2 (i.e., 2-to-1) then we are saying that A is twice as likely to occur as not. That is Pr(A)=2/3 and Pr(not A)=1/3.

Equivalence of binomial and Poisson approaches. . .

An odds interpretation. . .

Let's compare the expected occurrences in the $[2, 1]$ and $[2, 2]$ cells.

Within row 2, an occurrence is $\exp(\beta_2 + \beta_3)$ times as likely to be in column 2 than column 1.

That is, within row 2, the odds of being in column 2 (rather than column 1) is $\exp(\beta_2 + \beta_3) = \exp(\beta_2) \times \exp(\beta_3)$.

Note that the multiplicative change in column 2 odds between row 2 and row 1 is $\exp(\beta_3)$.

Attendance/Pass. . .

Equivalence of binomial and Poisson approaches. . .

Recall, our contingency table for this example is

```
> Freqs.df
      Attend Fail Pass
1 not.attend   27   19
2     attend   17   83
```

and the CI for $\exp(\beta_3)$ from the Poisson model is

```
> exp(confint(AP.pois))[4,]
Waiting for profiling to be done...
    2.5 %    97.5 %
 3.214049 15.552487
```

We now have the same interpretation from the Poisson model as from the binomial model – the odds of an attender (row 2) passing STATS 20x (column 2) are between 3.2 and 15.6 times the odds of a non-attender passing.

Section 16.5
Closing remarks

Equivalence of binomial and Poisson approaches. . .

Closing remarks

The binomial and Poisson models fitted above are so-called *saturated* models because there are as many parameters as there are observations, and hence zero degrees of freedom.

A saturated model has perfect fit. To see that, let's look at the fitted counts from the Poisson model fitted to `Freqs2.df`.

```
> predict(AP.pois,type="response")
 1  2  3  4
17 83 27 19
> Freqs2.df
      Attend Pass freq
1     attend fail   17
2     attend pass   83
3 not.attend fail   27
4 not.attend pass   19
```

The fitted counts are exactly the observed counts – a perfect fit!

Equivalence of binomial and Poisson approaches. . .

Closing remarks

- The odds interpretation that we are using for Poisson analysis of contingency tables is generally not appropriate for the other types of Poisson count data that were seen in Chapters 13 and 14.[6]

- The odds multiplier was $\exp(\beta_1)$ from the binomial model, and equivalently $\exp(\beta_3)$ from the Poisson model. This is commonly called the **odds-ratio**. E.g., in row 2 the odds-ratio (relative to row 1) for the column 2 outcome is between 3.2 and 15.6.

- In practice, the **Poisson approach is most widely used** because
  - The binomial approach is limited to a binary response category.
  - The Poisson approach can handle tables of arbitrary number of row and/or columns.
  - The Poisson approach can handle multi-way tables.
  - The Poisson approach can test other types of hypotheses, e.g., that all row outcomes are equally likely.

[6] You will lose marks if used inappropriately.

Section 16.6
Odds ratios

(This is an optional Section - your lecturer will advise if it is examinable)

Odds ratios

Recall, the contingency table for the attendance/pass example is

```
> Freqs.df
      Attend Fail Pass
1 not.attend   27   19
2     attend   17   83
```

and the fitted Poisson model is

```
> coef(summary(AP.pois))
                        Estimate Std. Error   z value     Pr(>|z|)
(Intercept)            3.2958369  0.1924501 17.125671 9.550242e-66
Attendattend          -0.4626235  0.3096136 -1.494197 1.351243e-01
Passpass              -0.3513979  0.2994472 -1.173489 2.405999e-01
Attendattend:Passpass  1.9370252  0.4006749  4.834407 1.335434e-06
```

Our estimated value of the odds-ratio is therefore

```
> exp(coef(AP.pois))[4]
Attendattend:Passpass
              6.93808
```

Odds ratios...

It can be shown that the estimated odds-ratio is simply the product of the diagonal values in the table, divided by the product of the off-diagonal values.

In this case

$$\widehat{\text{OR}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} = \frac{27 \times 83}{17 \times 19}$$

```
> options(digits=4) # Set the number of significant digits to 4
> OR=27*83/(17*19)
> OR
[1] 6.938
```

Inference for odds ratios

The distribution of the estimated odds ratio is difficult to obtain, and is skewed even in large samples.

However, the distribution of $\log(\widehat{\mathrm{OR}})$ is approximately normal for large samples.

It can be shown (see STATS 730) that the standard error of $\log(\widehat{\mathrm{OR}})$ is approximately

$$\mathrm{se}\left(\log(\widehat{\mathrm{OR}})\right) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

This can be used to obtain approximate confidence intervals for $\log(\widehat{\mathrm{OR}})$, which can be exponentiated to obtain CI's for $\widehat{\mathrm{OR}}$.

Numerous packages in `R` (e.g., `epitools`) will do these calculations for you.

Inference for odds ratios

Attendance/Pass example

Evaluating the above formula for the approximate standard error of $\log(\widehat{\mathrm{OR}})$ gives

```
> logOR.se=sqrt(1/17+1/83+1/27+1/19)
> logOR.se
[1] 0.4007
```

and the approximate 95% CI is therefore

```
> logOR.CI=log(OR) + c(-1,1)*1.96*logOR.se
> logOR.CI
[1] 1.152 2.722
```

The approximate 95% CI for the odds ratio is therefore

```
> exp(logOR.CI)
[1]  3.164 15.216
```

Compare with

```
> exp(confint.default(AP.pois)[4,] )
 2.5 % 97.5 %
 3.164 15.216
```

Section 16.7
Analysing Attendance/Pass using STATS 10x methods
1) difference in proportions
2) chi-squared test for association

(This is an optional Section - your lecturer will advise if it is examinable)

Method 1: Difference in proportions

We'll let p_1 and p_2 denote the probabilities of passing for attenders and non-attenders, respectively.

Recall that the standard error of a difference in proportions is estimated by

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

Our approximate 95% confidence interval for $p_1 - p_2$ is calculated as

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times se(\hat{p}_1 - \hat{p}_2)$$

where $\hat{p}_1$ and $\hat{p}_2$ are the proportions of passes. That is, $\hat{p}_1 = 83/100 = 0.83$ and $\hat{p}_2 = 19/46 = 0.413$.

In R:

```
> p1 = 83/100; p2 = 19/46
> se = sqrt(p1 * (1 - p1)/100 + p2 * (1 - p2)/46)
> ## Calculate the confidence interval
> p1 - p2 + 1.96 * c(-1, 1) * se
[1] 0.2567 0.5772
```

Attendance/Pass

Using the difference in proportions passing

The above 95% CI does not include 0, so this is a statistically significant difference (p-value < 0.05), so we can conclude that passing is highly associated with attendance.

This tell us that attenders have a probability of passing that is between 0.26 and 0.58 higher than that of non-attenders.

Method 2: Chi-square test for association in a contingency table

Recall:

```
> AP.df = read.table("Data/AttendPass.txt",header=T)
> AP.tbl=with(AP.df,table(Attend,Pass))
> AP.tbl
            Pass
Attend       fail pass
  attend       17   83
  not.attend   27   19
```

These counts are re-arranged in the form of a 2×2 contingency table.

Chi-squared goodness of fit test

Attendance/Pass...

The standard Chi-squared test is given by

```
> chisq.test(AP.tbl,correct=FALSE)

Pearson’s Chi-squared test

data:  AP.tbl
X-squared = 26, df = 1, p-value = 3e-07
```

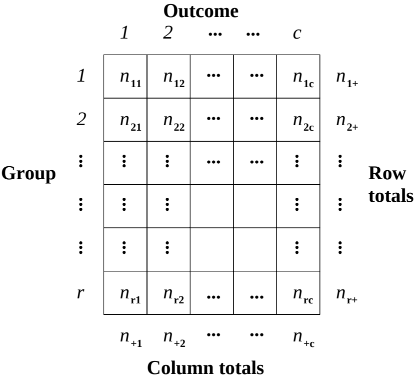
The null hypothesis of no association between attendance and course success. In this case, the test tell us that there is massive evidence against this hypothesis—which we already knew.

We will now examine how this test works.

Tests of association in a 2-way table

Notation

Here is a way of considering a general table of counts with an arbitrary number of levels for each factor variable.



Outcome

		1	2	⋯	⋯	c		
	1	n_{11}	n_{12}	⋯	⋯	n_{1c}	n_{1+}	
	2	n_{21}	n_{22}	⋯	⋯	n_{2c}	n_{2+}	
Group	⋮	⋮	⋮	⋯	⋯	⋮	⋮	**Row totals**
	⋮	⋮	⋮			⋮	⋮	
	⋮	⋮	⋮			⋮	⋮	
	r	n_{r1}	n_{r2}	⋯	⋯	n_{rc}	n_{r+}	
		n_{+1}	n_{+2}	⋯	⋯	n_{+c}		

Column totals

n_{ij} = number of subjects in group (row) i and having outcome (column) j.

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \text{total number of subjects.}$$

Tests of association in a 2-way table: χ^2 test

Deriving the χ^2 test using the Poisson assumption

The Chi[7]-squared (χ^2) test of independence (between row and column factors) compares the observed counts to those one would expect if there was no association between the row and column factors.

If the row and columns factors are independent then the probability of a subject being placed in cell (i, j) of the table equals the product of the probability that they are in row i (i.e., level i of the row factor occurs) times the probability that they are in column j (i.e., level j of the row factor occurs).

You have seen the χ^2 test before (I hope). Here, we are going to derive the χ^2 test using what we know about the properties of the Poisson distribution.

[7] Pronounced /kai/ - a long i sound like ride

Tests of association in a 2-way table: χ^2 test

Deriving the χ^2 test using the Poisson assumption...

- Our estimate of the row i probability is $\frac{n_{i+}}{n}$.
- Our estimate of the column j probability is $\frac{n_{+j}}{n}$.
- Under the null hypothesis of row and column independence, our estimate of the cell (i,j) probability is

$$\frac{n_{i+}}{n} \times \frac{n_{+j}}{n}$$

- So, under the null hypothesis, the expected value of the the number of subjects in cell (i,j) is estimated to be

$$\hat{n}_{ij} = n \times \frac{n_{i+}}{n} \times \frac{n_{+j}}{n}$$

Tests of association in a 2-way table: χ^2 test. . .

Deriving the χ^2 test using the Poisson assumption. . .

Next we have to determine whether the residuals, $n_{ij} - \hat{n}_{ij}$, are of sufficient magnitude to provide evidence against the null hypothesis.
We do this by calculating a Z-statistic for each cell:

$$Z_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{sd(n_{ij})}$$

where $sd(n_{ij})$ is the standard deviation of n_{ij}.

Now, we will utilize the assumption that the counts are Poisson distributed. Recall that Poisson count data have variance equal to the mean. We've estimated the mean to be $\hat{n}_{ij}$, so we estimate $sd(n_{ij})$ to be

$$sd(n_{ij}) = \sqrt{\mathrm{Var}(n_{ij})} = \sqrt{\mathrm{E}(n_{ij})} \approx \sqrt{\hat{n}_{ij}}$$

and so our Z-statistic for cell $[i, j]$ of the table is

$$Z_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}$$

Tests of association in a 2-way table: χ^2 test...

Deriving the χ^2 test using the Poisson assumption...

The chi-squared test statistic is given by summing the square of the Z_{ij} statistics over all cells in the table. That is,

$$X = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Under H_0 (no association between rows and columns), X has an approximate $\chi^2_{(r-1)(c-1)}$ distribution, where $(r-1)(c-1)$ is the degrees of freedom[8]. So, for a 2-by-2 table, it is approximately χ^2_1 distributed.

[8] Degrees of freedom are calculated as the number of independent data points minus the number of estimated parameters—can you do the maths here?

Tests of association in a 2-way table: χ^2 test…

The χ^2 test p-value

Large values of X provide evidence against H_0. The P-value is the "probability to the right of X", where the probability is from a $\chi^2_{(r-1)(c-1)}$ distribution.

In R, the probability that a χ^2_q is *less than* X is given by `pchisq(X,q)`. The P-value is given by 1 minus this amount. That is,

$$P\text{-value} = \texttt{1-pchisq(X,q)}$$

For a 2-by-2 table, we are working with χ^2_1, so the P-value is the probability that a value from a χ^2_1 distribution exceeds X. So, for a 2-by-2 table with $X = 3.1$, the P-value would be

```
> 1-pchisq(3.1,1)
[1] 0.07829
```

Tests of association in a 2-way table: χ^2 test…

Validity of the χ^2 test

NOTE: The χ^2 test uses the square of Z-statistics. These require that n_{ij} is at least approximately normally distributed.

Recall - this requires the expected counts to be sufficiently large.

That is why implementations of the χ^2 test often give warnings if some of the estimated expected counts ($\hat{n}_{ij}$) are “too small”. ‘Too small” is commonly taken to be < 5.

If this assumption is violated then there is a solution that uses a permutation technique called Fisher’s exact test.

Example—Attendance/Pass

```
> chisq.test(AP.tbl)

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

data:  AP.tbl
X-squared = 24, df = 1, p-value = 9e-07
```

This is a modified form of the chi-squared test that does a continuity correction to the Z_{ij} values, so as to correct for the fact that the cell counts are integer valued.

Attendance/Pass

Chi-squared goodness of fit test

The standard chi-squared test (with X defined as above) is given by

```
> chisq.test(AP.tbl,correct=FALSE)

Pearson’s Chi-squared test

data:  AP.tbl
X-squared = 26, df = 1, p-value = 3e-07
```

In this case, both forms of the test tell us that there is massive evidence against the null hypothesis of no association between attendance and course success.

If the difference between the corrected and standard chi-squared tests makes any meaningful difference then you are probably up to no good!

Attendance/Pass...

The expected values can be obtained by saving the result of chisq.test as an R object, and printing the $expected component of it.

```
> AP.chisq = chisq.test(AP.tbl, correct=FALSE)
> AP.chisq$expected
            Pass
Attend         fail  pass
  attend      30.14 69.86
  not.attend  13.86 32.14
```

The actual counts for comparison:

```
> AP.tbl
            Pass
Attend        fail pass
  attend        17   83
  not.attend    27   19
```