Associations

Introduction to Quantitative Ecology
Fall 2018
Chris Sutherland
csutherland@umass.edu

Group evaluations

1. Specifically for the assignment due tomorrow, how would you describe your comfort levels with R and Excel?

   A) Uncomfortable with both
   B) Comfortable with Excel, but not R
   C) Comfortable with R, but not Excel
   D) Comfortable with both

Group evaluations

Is there a significant association between salamander *sex* and *habitat type*?

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

1. Calculate χ^2.
2. Calculate DF
3. Is there a significant association?

	Dry	Moist	Wet	$\sum$ Row
Female	*370*	*198*	*187*	
Male	*359*	*110*	*160*	
$\sum$ Col				

df	0.05
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

Associations

What are associations?

- dealing with two categorical variables
- known as a *contingency table*
  - data are *cross tabulated* frequencies
  - each cell represents a count

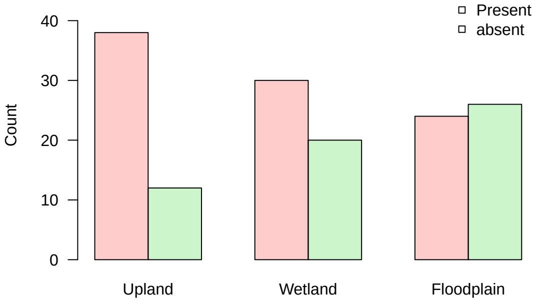
Associations

- dealing with two categorical variables
- known as a *contingency table*
  - data are *cross tabulated* frequencies
  - each cell represents a count

Salamander	Upland	Wetland	Floodplain
Present	*38*	*30*	*24*
Absent	*12*	*20*	*26*

Associations

- dealing with two categorical variables
- known as a *contingency table*
  - data are *cross tabulated* frequencies
  - each cell represents a count
- visualize using a *bar chart*



Associations

The key question with contingency tables:

- is there a significant association between the categorical variable A and categorical variable B?
- are these numbers different to what we would expect to see by chance?

Associations

The key question with contingency tables:

- is there a significant association between the categorical variable A and categorical variable B?
- are these number different to what we would expect to see by chance?
- answer using the *Chi-squared test*

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

- O: the observed data
- E: the expected value if there was *no association*
- χ^2: the test statistic

Associations

In a Chi-squared test from the following contingency table, I get a test statistic of $\chi^2 = 8.32$. Is there a significant association between spotted salamander presence and vernal pool type at the 5% level?

Salamander	Upland	Wetland	Floodplain
Present	*38*	*30*	*24*
Absent	*12*	*20*	*26*

A) Yes, its significant!

B) No, it's not significant!

df	0.05
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

Associations

In a Chi-squared test from the following contingency table, I get a test statistic of $\chi^2 = 8.32$. Is there a significant association between spotted salamander presence and vernal pool type at the 5% level?

Salamander	Upland	Wetland	Floodplain
Present	*38*	*30*	*24*
Absent	*12*	*20*	*26*

A) **Yes, its significant!**

B) No, it's not significant!

df	0.05
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

Associations - observed

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

The observed values (O):

- the data we observe (obviously!)
- cross tabulated counts

Associations - expected

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

The expected values (E):

- the data we would expect by change
- the data we would expect if there was no association

$$E = \frac{\text{row total} \cdot \text{col total}}{\text{grand total}}$$

Associations - expected

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

Degrees of freedom:

$$DF = (no.columns - 1) \times (no.rows - 1)$$

Associations - expected

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

Table I. Critical Values of χ^2

	Level of Significance for Two-tailed Test					
df	.20	.10	.05	.02	.01	.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.60	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.80	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.00	28.26	30.58	37.70

Associations - invertebrate example

Invertebrate group habitat selection:

	Ant	Bug	Beetle	Total
Upper leaf	15	13	68	96
Lower leaf	12	11	15	38
Stem	65	78	5	148
Bud	3	21	3	27
Total	95	123	91	309

Associations - invertebrate example

First we need to state the hypotheses!

- Null hypothesis:

?

Associations - invertebrate example

First we need to state the hypotheses!

- Null hypothesis:
- the *no habitat preference* hypothesis
- random!

“There is no association between invertebrate group and habitat”

Associations - invertebrate example

First we need to state the hypotheses!

- Null hypothesis:
- the *no habitat preference* hypothesis
- random!

"There is no association between invertebrate group and habitat"

- Alternative hypothesis:

?

Associations - invertebrate example

First we need to state the hypotheses!

- Null hypothesis:
- the *no habitat preference* hypothesis
- random!

“There is no association between invertebrate group and habitat”

- Alternative hypothesis:
- the *habitat preference* hypothesis
  - direction not explicitly stated
  - can be positive or negative
- not random!

“There *is* an association between invertebrate group and habitat”

Associations - invertebrate example

Demo in `Excel`

Associations - invertebrate example

- The test statistic is $\chi^2 = 146.98$.
- What do we conclude?

Associations - invertebrate example

- The test statistic is $\chi^2 = 146.98$.
- What do we conclude?
  - $p < 0.05$
  - reject the null hypothesis
  - accept the alternative hypothesis
  - there *is* a significant association between inverts and habitat!

Associations - invertebrate example

- The test statistic is $\chi^2 = 146.98$.
- What do we conclude?
  - $p < 0.05$
  - reject the null hypothesis
  - accept the alternative hypothesis
  - there *is* a significant association between inverts and habitat!
- BUT! which associations are significant?

	Ant	Bug	Beetle	Total
Upper leaf	15	13	68	96
Lower leaf	12	11	15	38
Stem	65	78	5	148
Bud	3	21	3	27
Total	95	123	91	309

Associations - significant associations

Two ways we can evaluate which associations are likely to be significant:

1. Cell-specific χ^2 values
   - greater than 3.8 is likely to be significant
   - 3.8 is significant test statistic with 1 degree of freedom

Associations - significant associations

Two ways we can evaluate which associations are likely to be significant:

1. Cell-specific χ^2 values
   - greater than 3.8 is likely to be significant
   - 3.8 is significant test statistic with 1 degree of freedom
2. Pearson residuals
   - provides sign of association
   - provides relative size of the association
   - if residual is >2 or < -2 then likely to be significant

$$Residual = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

Associations - invertebrate example

Back to `Excel`

Analyzing contingency table data in `Excel`

Two `Excel` functions for doing Chi-squared or Goodness of fit tests but no method in *Analysis Tool Pack*:

Analyzing contingency table data in `Excel`

Two `Excel` functions for doing Chi-squared or Goodness of fit tests but no method in *Analysis Tool Pack*:

- `CHITEST(`*observed*,*expected*`)`
  - must calculate the *expected* values
  - must be in the same table format

Analyzing contingency table data in `Excel`

Two `Excel` functions for doing Chi-squared or Goodness of fit tests but no method in *Analysis Tool Pack*:

- `CHITEST(`*observed*,*expected*`)`
  - must calculate the *expected* values
  - must be in the same table format
- `CHIDIST(`*Chi value*,*derees of freedom*`)`
  - must calculate χ^2 and DF
  - just calculates a p-value

Demo in `Excel`

Analyzing contingency table data in R

- Chi-square test in R
  - `chisq.test(`*contingency table*`)`
  - data must be formatted like a contingency table
  - can make a `data.frame`

```
# a dataframe
tab.df <- data.frame(Ant = c(15,12,65,3),
                     Bug = c(13,11,78,21),
                     Beetle = c(68,15,5,3))
rownames(tab.df) <- c("Upper", "Lower", "Stem", "Bud")
tab.df
      Ant Bug Beetle
Upper  15  13     68
Lower  12  11     15
Stem   65  78      5
Bud     3  21      3
```

Analyzing contingency table data in R

- Chi-square test in R
  - `chisq.test(`*contingency table*`)`
  - data must be formatted like a contingency table
  - can make a `matrix`

```
# a dataframe
tab.mat <- matrix(c(15,12,65,3,
                    13,11,78,21,
                    68,15,5,3), nrow=4, ncol=3, byrow=FALSE)
rownames(tab.mat) <- c("Upper", "Lower", "Stem", "Bud")
colnames(tab.mat) <- c("Ant", "Bug", "Beetle")
tab.mat
      Ant Bug Beetle
Upper  15  13     68
Lower  12  11     15
Stem   65  78      5
Bud     3  21      3
```

Analyzing contingency table data in Excel

- Chi-square test in R
  - `chisq.test(`*contingency table*`)`
  - data must be formatted like a contingency table
  - can make a `data.frame` or a `matrix`

```
# conduct the Chi-square test
chisq.test(tab.df)

    Pearson's Chi-squared test

data:  tab.df
X-squared = 146.98, df = 6, p-value < 2.2e-16
```

Analyzing contingency table data in Excel

- Chi-square test in R
  - `chisq.test(`*contingency table*`)`
  - data must be formatted like a contingency table
  - can make a `data.frame` or a `matrix`

```
# extract the expected values
chisq.test(tab.df)$expected
            Ant      Bug    Beetle
Upper 29.514563 38.21359 28.271845
Lower 11.682848 15.12621 11.190939
Stem  45.501618 58.91262 43.585761
Bud    8.300971 10.74757  7.951456
```

Analyzing contingency table data in `Excel`

- Chi-square test in `R`
  - `chisq.test(`*contingency table*`)`
  - data must be formatted like a contingency table
  - can make a `data.frame` or a `matrix`

```
# calculate the Pearson residuals
obs <- tab.df
exp <- chisq.test(tab.df)$expected
(obs-exp) / sqrt(exp)
             Ant       Bug    Beetle
Upper -2.6716883 -4.078738  7.471733
Lower  0.0927883 -1.060930  1.138636
Stem   2.8905810  2.486807 -5.844599
Bud   -1.8398862  3.127314 -1.755940
```

Analyzing contingency table data in Excel

What can we conclude from our analysis of the invertebrate data using the Chi-square test for association?

```
# conduct the Chi-square test
chisq.test(tab.df)

	Pearson's Chi-squared test

data:  tab.df
X-squared = 146.98, df = 6, p-value < 2.2e-16
```

```
# Pearson residuals
(obs-exp) / sqrt(exp)
             Ant       Bug    Beetle
Upper -2.6716883 -4.078738  7.471733
Lower  0.0927883 -1.060930  1.138636
Stem   2.8905810  2.486807 -5.844599
Bud   -1.8398862  3.127314 -1.755940
```

‘Pioneer Valley Camera Trapping Data’

The data:

- 142 camera traps placed throughout the valley
- each camera has 2 categorical covariates:
  - land use: ‘altered’ (A), ‘natural’ (N) and ‘urban’ (U)
  - scent lure: ‘Badlands Bob’ (BB) and ‘Powder River’ (PR)
- we will focus on four species:
  - bobcat *lynx rufus*
  - domestic cat *felis catus*
  - coyote *canis latrans*
  - domestic dog *canis familiaris*

Group exercise

Using the ‘Pioneer Valley Camera Trapping Data’ investigate whether there is a statistically significant association between the four species and:

1. habitat type
2. scent lure

(i.e., conduct two analyses)