Categorical: Contingency tables

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this lecture

- Extend 2 x 2 contingency tables to larger tables
  - More variable categories: 2×3 and larger
  - More variables: $2 \times 2 \times 2$ tables
- Chi-square tests for these tables

- Probing the tables
- Residuals

2 More variable categories

2.1 $I \times 2$ and $2 \times J$ tables

2.1.1 2×2 tables and beyond...

- We've only looked at 2×2 tables
  - Extend these tables in terms of rows or columns
    - Just more rows: $I \times 2$
    - Just more columns: $2 \times J$
    - More rows and more columns: $I \times J$

2.1.2 `Whickam2` data with a twist

```
   Outcome Smoker Age AgeGroup Alive
1    Alive    Yes  23    18-64     1
2    Alive    Yes  18    18-64     1
3     Dead    Yes  71      65+     0
4    Alive     No  67      65+     1
5    Alive     No  64    18-64     1
6    Alive    Yes  38    18-64     1
7    Alive    Yes  45    18-64     1
8     Dead     No  76      65+     0
9    Alive     No  28    18-64     1
10   Alive     No  27    18-64     1
```

- Split the `Age` variable into **3 groups**
  - 18 to 40
  - 41 to 64
  - 65+

2.1.3 Convert `Age` into 3 categories

```
Whickham2 <- Whickham2 %>%
    mutate(AgeGroup3 = ifelse(Age %in% 18:40, 1,
           ifelse(Age %in% 41:64, 2, 3)))
head(Whickham2)
```

```
  Outcome Smoker Age AgeGroup Alive AgeGroup3
1   Alive    Yes  23    18-64     1         1
2   Alive    Yes  18    18-64     1         1
3    Dead    Yes  71      65+     0         3
4   Alive     No  67      65+     1         3
5   Alive     No  64    18-64     1         2
6   Alive    Yes  38    18-64     1         1
```

2.1.4 `Agegroup3` versus `Alive`: Observed

	Dead	Alive	Sum
18 to 40	19	521	540
41 to 64	141	390	531
65+	209	34	243
Sum	369	945	1314

2.1.5 `Agegroup3` versus `Alive`: Expected

	Dead	Alive	Sum
18 to 40	151.644	388.356	540
41 to 64	149.116	381.884	531
65+	68.240	174.760	243
Sum	369.000	945.000	1314

2.1.6 `Agegroup3` versus `Alive`: Chi-square

$\chi^2 = \sum \left(\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} \right) = \sum \left(\frac{(O-E)^2}{E} \right) =$

$\frac{(19-151.644)^2}{151.644} + \frac{(141-149.116)^2}{149.116} + \frac{(209-68.24)^2}{68.24} + \frac{(521-388.356)^2}{388.356} + \frac{(390-381.884)^2}{381.884} + \frac{(34-174.76)^2}{174.76} =$

$116.024 + 0.442 + 290.351 + 45.305 + 0.173 + 113.375 = 565.669$

2.1.7 `Agegroup3` versus `Alive`: Chi-square

- Degrees of freedom $= (I - 1) \times (J - 1) = (3 - 1) \times (2 - 1) = 2$
  - $\chi^2_{critical}(2) = 5.99$
  - $565.669 > 5.99$
  - Reject H_0 that `AgeGroup3` and `Alive` are independent
- But then what?
  - What is different?
  - Similar to ANOVA
    - With 3 groups (levels), which one(s) are different from each other?

2.2 Partitioning χ^2

2.2.1 Partitioned chi-square

- Chi-square statistics can be split up (partitioned):
  - `Alive` by (`18 to 40` vs `65+`)
  - `Alive` by ((`18 to 40` and `41 to 64`) vs `65+`)
  - These are two *independent* tests
- Chi-square for all *independent* tests add up to chi-square for complete table
  - Kind of
- Degrees of freedom also add up

2.2.2 Orthogonal tests

- Orthogonal partitioning of a contingency table is similar to coding orthogonal contrasts for ANOVA

Orthogonal partitioning			
`Alive` by (`18 to 40` vs `41 to 64`)	+1	-1	0
`Alive` by ((`18 to 40` and `41 to 64`) vs `65+`)	-0.5	-0.5	+1

Not orthogonal partitioning			
`Alive` by (`18 to 40` vs `65+`)	+1	0	-1
`Alive` by (`41 to 64` vs `65+`)	0	+1	-1

2.2.3 `Agegroup3` versus `Alive`: Partition 1

- Observed:

	Dead	Alive	Sum
18 to 40	19	521	540
41 to 64	141	390	531
Sum	160	911	1071

- Expected:

	Dead	Alive	Sum
18 to 40	80.67	459.33	540
41 to 64	79.33	451.67	531
Sum	160.00	911.00	1071

```
    Pearson's Chi-squared test

data:  Age3_Alive[c(1, 2), ]
X-squared = 111.79, df = 1, p-value < 0.00000000000000022
```

2.2.4 `Agegroup3` versus `Alive`: Partition 2

- Observed:

	Dead	Alive	Sum
18 to 64	160	911	1071
65+	209	34	243
Sum	369	945	1314

- Expected:

	Dead	Alive	Sum
18 to 64	300.76	770.24	1071
65+	68.24	174.76	243
Sum	369.00	945.00	1314

```
    Pearson's Chi-squared test

data:  Age_Alive
X-squared = 495.33, df = 1, p-value < 0.00000000000000022
```

2.2.5 Partitioned chi-square

- Overall: $\chi^2(2) = 565.67$
- Partition 1: $\chi^2(1) = 111.79$
- Partition 2: $\chi^2(1) = 495.33$
- $111.79 + 495.33 = 607.12 \approx 565.67$
  - For the χ^2 statistic, the sum will be approximate
    - Closer for larger *samples* and larger *tables*
  - A slightly different statistic, G^2, will always sum perfectly
    - G^2 can also be partitioned in the same

2.2.6 G^2 statistic

- $G^2 = 2\Sigma\left(n_{ij} \times ln\left(\frac{n_{ij}}{\mu_{ij}}\right)\right)$
  - Also called "likelihood ratio test statistic"
  - Compare to χ^2 distribution with $(I-1) \times (J-1)$ df

2.2.7 G^2 statistic

- Overall 3×2 table

```
    Log likelihood ratio (G-test) test of independence without correction

data:  Age3_Alive
G = 584.41, X-squared df = 2, p-value < 0.00000000000000022
```

- Just `18 to 40` vs `41 to 64`

```
    Log likelihood ratio (G-test) test of independence without correction

data:  Age3_Alive[c(1, 2), ]
G = 124.02, X-squared df = 1, p-value < 0.00000000000000022
```

- Combined (`18 to 40` and `41 to 64`) vs `65+`

```
    Log likelihood ratio (G-test) test of independence without correction

data:  Age_Alive
G = 460.39, X-squared df = 1, p-value < 0.00000000000000022
```

2.3 Residuals

2.3.1 Residuals

- Residuals exist for χ^2 just like linear regression
- **Raw residual** = observed - expected = $n_{ij} - \hat{\mu}_{ij}$
- **Standardized residual** divides by std error of raw residuals
  - $\frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$
  - where $\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}$ is std error of raw residuals under H_0
  - and $p_{i+} = n_{i+}/n$ and $p_{+j} = n_{+j}/n$

2.3.2 Observed and expected frequencies

- Observed

	Dead	Alive
18 to 40	19	521
41 to 64	141	390
65+	209	34

- Expected

	Dead	Alive
18 to 40	151.644	388.356
41 to 64	149.116	381.884
65+	68.240	174.760

2.3.3 Residuals

- Raw residuals

	Dead	Alive
18 to 40	-132.644	132.644
41 to 64	-8.116	8.116
65+	140.760	-140.760

- Standardized residuals

	Dead	Alive
18 to 40	-16.549	16.549
41 to 64	-1.015	1.015
65+	22.256	-22.256

2.3.4 Standardized residuals

- Under H_0, variables are **independent**
  - Observed cell frequencies = expected cell frequencies
    - Residuals tell you how much each cell deviates from this
    - Large standardized residual = cell shows lack of fit from H_0
- Standardized ≈ normal distribution
  - Expect about 5% of residuals to be greater than ±2
    - Look at standardized residual greater than ±2
    - In small tables, this is way off

2.3.5 Standardized residuals

- Observed

	Dead	Alive
18 to 40	19	521
41 to 64	141	390
65+	209	34

- Standardized residuals

	Dead	Alive
18 to 40	-16.549	16.549
41 to 64	-1.015	1.015
65+	22.256	-22.256

3 More variables

3.1 Conditional and marginal effects

3.1.1 Adding a 3rd variable

- **Control for** a potentially confounding third variable, like **smoking**

- Relationship between `AgeGroup` (X) and `Alive` (Y)
    - What if **smokers** have one relationship between X and Y
    - But **non-smokers** have a different relationship between X and Y?
- From last time: `Smoker` and `Alive` had an unexpected pattern
    - Smokers were **less likely to die** than non-smokers?
    - What if smokers are younger than non-smokers and that's what's really going on?

3.1.2 Adding a third variable: Z = `Smoker`

AgeGroup	Alive	Smoker	Freq
18 to 64	Dead	No	65
65+	Dead	No	165
18 to 64	Alive	No	474
65+	Alive	No	28
18 to 64	Dead	Yes	95
65+	Dead	Yes	44
18 to 64	Alive	Yes	437
65+	Alive	Yes	6

3.1.3 3 variables = 3-way = 3D

- 2 variables = 2-way or 2D table
    - 3 variables = 3-way or 3D table
- Two ways to look at a 3D table
    - **Partial table** (a.k.a. conditional table)
    - **Marginal table**

3.1.4 Partial tables

- Slice 3D table into more 2D tables
    - 2-way table of X vs Y for each level of Z
- *Conditional* on levels of Z
    - Remove effect of Z by holding it constant at specific levels
- *Conditional associations*
    - e.g., conditional χ^2

```
, , Smoker = No

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64   65   474  539
  65+       165    28  193
  Sum       230   502  732

, , Smoker = Yes

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64   95   437  532
  65+        44     6   50
  Sum       139   443  582

, , Smoker = Sum

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64  160   911 1071
  65+       209    34  243
  Sum       369   945 1314
```

3.1.5 Marginal table

- 2D table ignoring Z
  - 2-way table of X vs Y
- Collapse across levels of Z
  - Add up across
  - No information about Z
- *Marginal associations*

```
          Alive
AgeGroup   Dead Alive  Sum
  18 to 64  160   911 1071
  65+       209    34  243
  Sum       369   945 1314
```

3.1.6 `Alive` vs `AgeGroup`: Conditional on `Smoker`

- Frequencies

```
, , Smoker = No

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64   65   474  539
  65+       165    28  193
  Sum       230   502  732

, , Smoker = Yes

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64   95   437  532
  65+        44     6   50
  Sum       139   443  582

, , Smoker = Sum

          Alive
AgeGroup   Dead Alive  Sum
  18 to 64  160   911 1071
  65+       209    34  243
  Sum       369   945 1314
```

- Proportions

```
# A tibble: 4 x 6
# Groups:   AgeGroup, Smoker [4]
  AgeGroup Smoker Alive count  prop totaln
  <fct>    <fct>  <int> <int> <dbl>  <int>
1 18-64    No         1   474 0.879    539
2 18-64    Yes        1   437 0.821    532
3 65+      No         1    28 0.145    193
4 65+      Yes        1     6 0.12      50
```

3.1.7 `Alive` vs `Smoker`: Conditional on `AgeGroup`

- Frequencies

```
, , AgeGroup = 18-64

      Alive
Smoker Dead Alive  Sum
   No    65   474  539
   Yes   95   437  532
   Sum  160   911 1071

, , AgeGroup = 65+

      Alive
Smoker Dead Alive  Sum
   No   165    28  193
   Yes   44     6   50
   Sum  209    34  243

, , AgeGroup = Sum

      Alive
Smoker Dead Alive  Sum
   No   230   502  732
   Yes  139   443  582
   Sum  369   945 1314
```

- Proportions

```
# A tibble: 4 x 5
# Groups:   Smoker, AgeGroup [4]
  Smoker AgeGroup Alive count  prop
  <fct>  <fct>    <int> <int> <dbl>
1 No     18-64        1   474 0.879
2 No     65+          1    28 0.145
3 Yes    18-64        1   437 0.821
4 Yes    65+          1     6 0.12
```

3.1.8 `Alive` vs `AgeGroup`: Ignoring (marginal on) `Smoker`

- Frequencies

```
          Alive
AgeGroup   Dead Alive  Sum
  18 to 64  160   911 1071
```

```
  65+        209    34  243
  Sum        369   945 1314
```

- Proportions

```
# A tibble: 2 x 4
# Groups:   AgeGroup [2]
  AgeGroup Alive count  prop
  <fct>    <int> <int> <dbl>
1 18-64        1   911 0.851
2 65+          1    34 0.140
```

3.1.9 `Alive` vs `Smoker`: Ignoring (marginal on) `AgeGroup`

- Frequencies

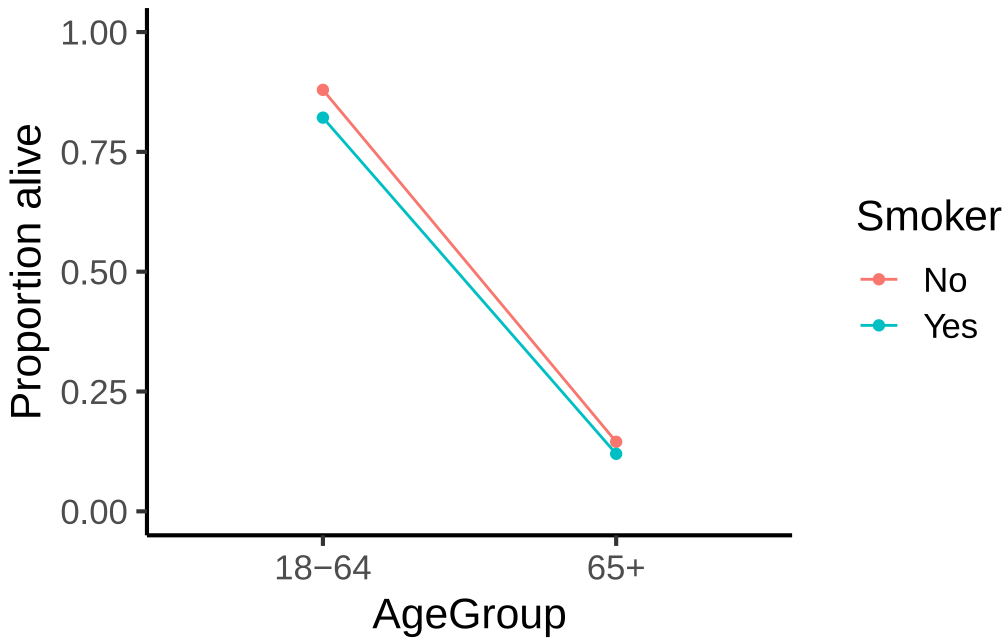
```
      Alive
Smoker Dead Alive  Sum
   No   230   502  732
   Yes  139   443  582
   Sum  369   945 1314
```

- Proportions

```
# A tibble: 2 x 5
# Groups:   Smoker [2]
  Smoker Alive count  prop totaln
  <fct>  <int> <int> <dbl>  <int>
1 No         1   502 0.686    732
2 Yes        1   443 0.761    582
```

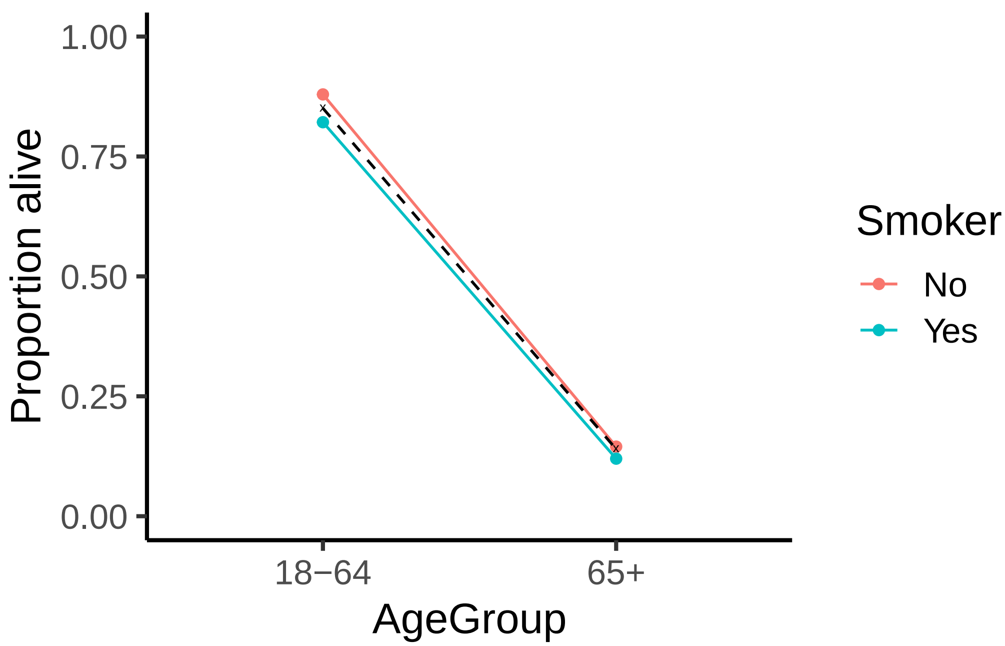
3.1.10 Plot: Proportion `Alive` by `AgeGroup`, by `Smoker`

- **Conditional** effect of `AgeGroup`, conditional on `Smoker`…



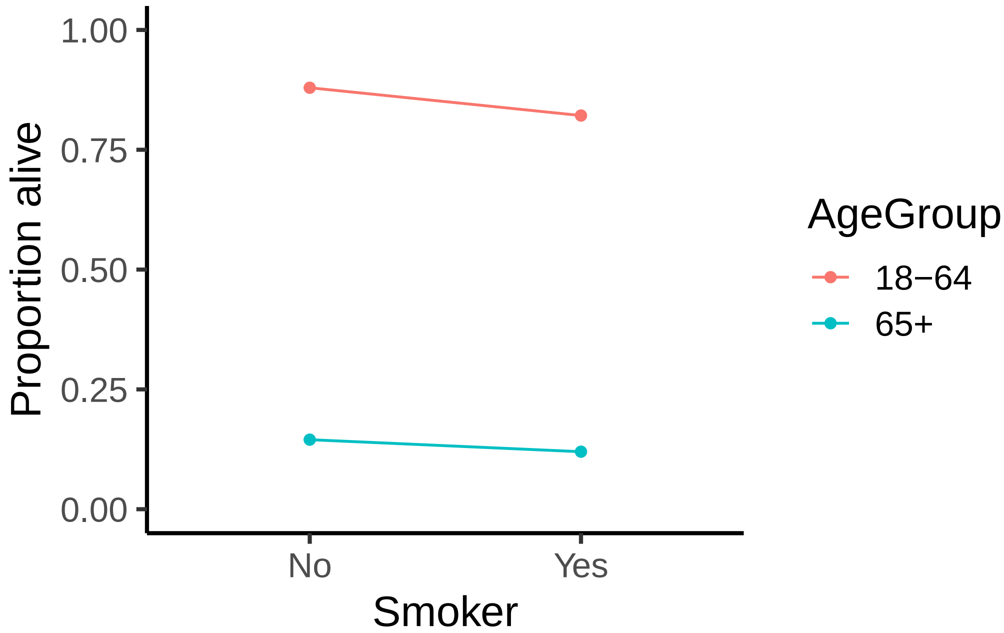
3.1.11 Plot: Proportion `Alive` by `AgeGroup`, by `Smoker`

- …plus **marginal effect**, ignoring `Smoker`



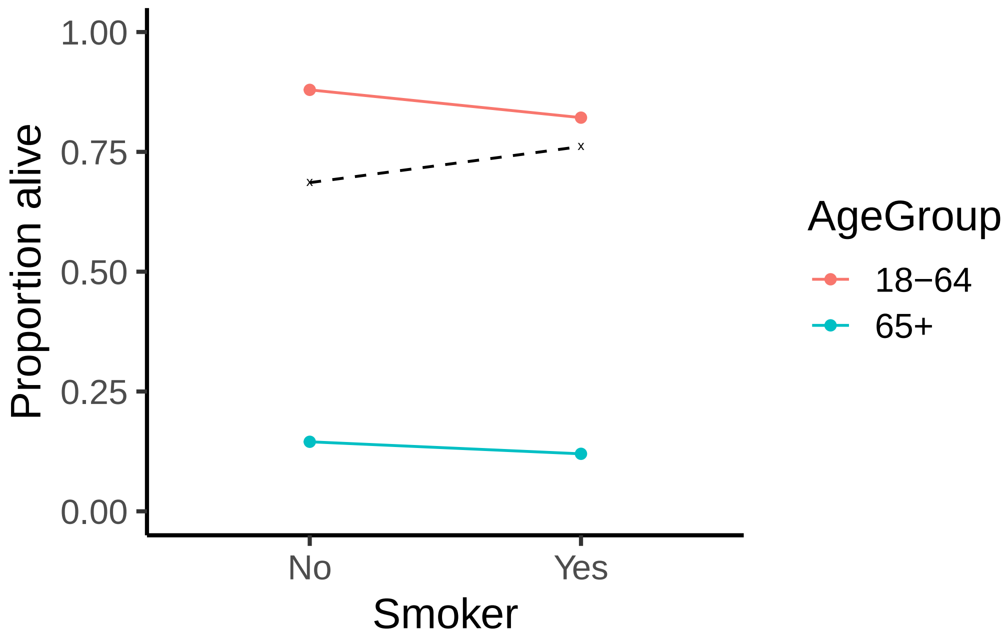
3.1.12 Plot: Proportion `Alive` by `Smoker`, by `AgeGroup`

- **Conditional** effect of `Smoker`, conditional on `AgeGroup`...



3.1.13 Plot: Proportion `Alive` by `Smoker`, by `AgeGroup`

- ...plus **marginal effect**, ignoring `AgeGroup`



3.2 Simpson's paradox

3.2.1 Conditional vs marginal effects

- **Conditional effects** take Z into account
  - Smokers **less likely to be alive** than non-smokers among *young people*
  - Smokers **less likely to be alive** than non-smokers among *old people*
- **Marginal effects** ignore Z
  - Smokers **more likely to be alive** than non-smokers

3.2.2 Simpson's paradox

- Opposite direction *marginal* vs *conditional* effects
- Not unique to contingency tables
  - Can happen any time you *ignore a confounder*
  - Change in direction of effect after adding covariate
- Related to
  - Lord's paradox: ANCOVA versus difference scores
  - Suppression effects: Change relationship by adding covariate
  - Ecological fallacy: Opposite direction effects at higher vs lower level

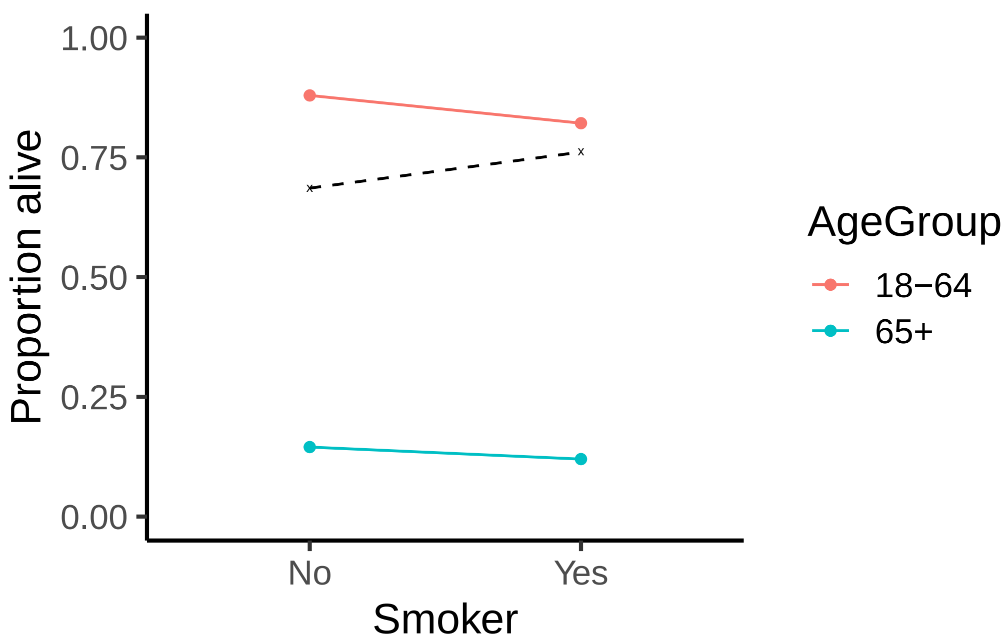
3.2.3 Why does it happen?

- Relationship between the confounder and other variables besides Y
  - In this case, the relationship between `AgeGroup` and `Smoker`

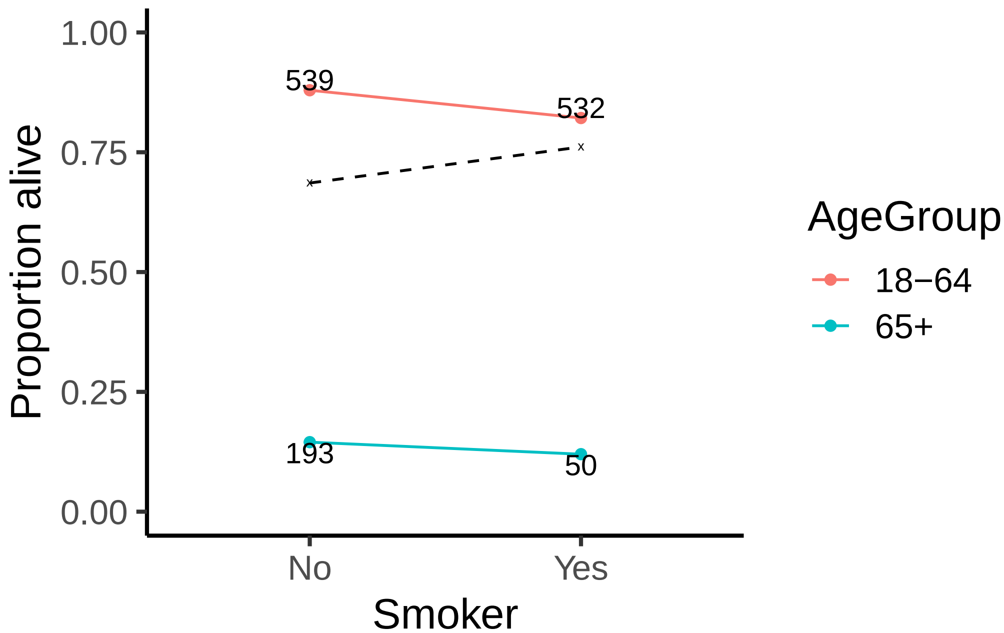
```
          Smoker
AgeGroup     No  Yes  Sum
  18 to 64  539  532 1071
  65+       193   50  243
  Sum       732  582 1314
```

- Younger people are $\frac{532/539}{50/193} = \frac{0.987}{0.259} = 3.81$ times more likely to be smokers than older people
  - $1071/243 = 4.41$ times more young people than older people
  - A lot of young people, who are more likely to smoke and less likely to die

3.2.4 Plot: Simpson's paradox



3.2.5 Plot: Total n for each proportion



3.2.6 Simpson's paradox

- Not really a paradox
  - Just a different way of looking at the information
- When you have 3 variables
  - Focus is XY relationship
  - But also pay attention to the ZY and XZ relationships

3.3 Conditional and marginal odds ratios

3.3.1 Conditional and marginal effects

- Partial tables
  - Conditional associations odds ratios, chi-square
  - Association between X and Y, at a given value of Z
- Marginal tables
  - Marginal associations, odds ratios, chi-square
  - Association between X and Y, ignoring Z

3.3.2 Partial tables

```
, , AgeGroup = 18-64

      Alive
Smoker Dead Alive  Sum
   No    65   474  539
   Yes   95   437  532
   Sum  160   911 1071

, , AgeGroup = 65+

      Alive
Smoker Dead Alive  Sum
   No   165    28  193
   Yes   44     6   50
   Sum  209    34  243

, , AgeGroup = Sum

      Alive
Smoker Dead Alive  Sum
   No   230   502  732
   Yes  139   443  582
   Sum  369   945 1314
```

3.3.3 Conditional odds ratios: Young people

- $\hat{\theta} = \frac{474/65}{437/95} = \frac{7.292}{4.6} = 1.585$
- Odds of *non-smoker* being alive = 7.292
  - A non-smoker is 7.292 times more likely to be alive than dead
- Odds of *smoker* being alive = 4.6
  - A smoker is 4.6 times more likely to be alive than dead
- Odds ratio = 1.585: Odds of a *non-smoker* being alive is 1.585 times the odds of an *smoker* being alive
  - Non-smokers are more likely to be alive than smokers

3.3.4 Conditional odds ratios: Older people

- $\hat{\theta} = \frac{28/165}{6/44} = \frac{0.170}{0.136} = 1.248$
- Odds of *non-smoker* being alive = 0.170
    - A non-smoker is 0.170 times more likely to be alive than dead
- Odds of *smoker* being alive = 0.136
    - A smoker is 0.136 times more likley to be alive than dead
- Odds ratio = 1.248: Odds of a *non-smoker* being alive is 1.248 times the odds of an *smoker* being alive
    - Non-smokers are more likely to be alive than smokers

3.3.5 Marginal table

```
      Alive
Smoker Dead Alive  Sum
   No   230   502  732
   Yes  139   443  582
   Sum  369   945 1314
```

3.3.6 Marginal odds ratio

- $\hat{\theta} = \frac{502/230}{443/139} = \frac{2.183}{3.187} = 0.685$
- Odds of *non-smoker* being alive = 2.183
    - A non-smoker is 2.183 times more likely to be alive than dead
- Odds of *smoker* being alive = 3.187
    - A smoker is 3.187 times more likley to be alive than dead
- Odds ratio = 0.685: Odds of a *non-smoker* being alive is 0.685 times the odds of an *smoker* being alive
    - Non-smokers are **less likely to be alive** than smokers

3.3.7 Conditional and marginal effects

- Neither tell the whole story usually
  - Look at both
- We looked at odds ratios
  - Also consider conditional and marginal difference in proportion and relative risk, if applicable

3.4 Conditional and marginal independence

3.4.1 Conditional vs marginal independence

- Variables can be
  - Conditionally independent
  - Marginally independent
  - Both
  - Neither
- Depends on XY relationship, as well as ZY and XZ relationships

3.4.2 Conditional independence...

- Clinic 1

	Success	Failure	
Treatment A	18	12	30
Treatment B	12	8	20
	30	20	50

- Odds of success is 1.5 times odds of failure for both treatments
- OR $= \frac{18/12}{12/8} = \frac{1.5}{1.5} = 1$

- Clinic 2

	Success	Failure	
Treatment A	2	8	10
Treatment B	8	32	40
	10	40	50

- Odds of failure is 4 times odds of success for both treatments
- OR $= \frac{2/8}{8/32} = \frac{.25}{.25} = 1$

3.4.3 ... but not marginal independence

- Combine both clinics into a single large sample

	Success	Failure	
Treatment A	20	20	40
Treatment B	20	40	60
	40	60	100

- Odds of success for treatment A is twice odds of success for treatment B
- OR $= \frac{20/20}{20/40} = \frac{1}{0.5} = 2$

3.4.4 What to do?

- Look at both **partial** and **marginal** tables
- Unless XY relationship is same at every level of Z, **partial and marginal effects will tell different stories**
- Very similar to covariates in ANOVA or regression
  - Main effect for XY relationship doesn't tell you the complete story if
    - Covariate is related to the outcome
    - Covariate and predictor are related
    - Interaction of covariate and predictor

4 Summary

4.1 Summary

4.1.1 Summary of this week

- Extend to **more levels** and/or **more variables**
  - Additional complexity with
    - Partitioning effects
    - Partial vs marginal effects

4.1.2 Summary of this section

- Contingency tables
  - 2×2 and larger
- Measures of relationship
  - Difference in proportion
  - Relative risk
  - Odds ratio
  - Chi-square tests
- Extend these to larger tables