Lab 01C: Week 4

Contents

The **specific aims** of this lab are:

- to practice statistical thinking with categorical and cross tabulated data
- develop understanding of chi-squared tests for homogeneity and independence
- become familiar with using Monte Carlo simulation in a contingency table context
- to generate different bar plots highlighting different features.

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO2 Extract and combine data from multiple data resources.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Quick quiz

1.1

An appropriate test to see if there is an association between hair colour (black, brown, blonde, red) and the presence of male-pattern baldness (none, moderate, severe) is:

a. Chi-squared goodness of fit test

b. Chi-squared test of independence

c. Test if the correlation coefficient is significantly different to zero

d. Check if the CI for the log odds ratio contains 1

1.2

In a test to see if there is an association between hair colour (black, brown, blonde, red) and the presence of male-pattern baldness (none, moderate, severe), the appropriate test statistic follows what type of distribution?

a. chi-squared with 3 degrees of freedom χ^2_3

b. chi-squared with 4 degrees of freedom χ^2_4

c. chi-squared with 6 degrees of freedom χ^2_6

d. chi-squared with 7 degrees of freedom χ^2_7

e. chi-squared with 12 degrees of freedom χ^2_{12}

1.3 TV violence

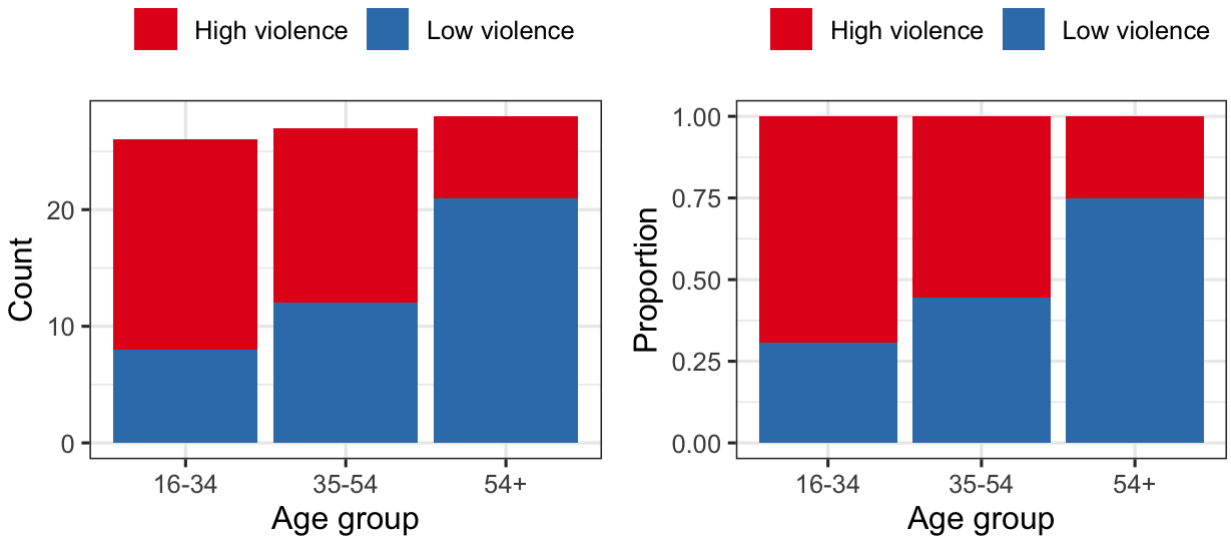
A study of the amount of violence viewed on television as it relates to the age of the viewer yields the results shown in the accompanying table for 81 people.

	Age		
Viewing	$16-34$	$35-54$	55 and over
Low violence	8	12	21
High violence	18	15	7

Does it look like there's a significant relationship between age group and violence viewing preference? No need to do a test at this point, just consider the numbers, and the visualisations below.

```
library("tidyverse")
```

```r
x = matrix(c(8, 18, 12, 15, 21, 7), ncol = 3)
colnames(x) = c("16-34", "35-54", "54+")
rownames(x) = c("Low violence", "High violence")
y = x %>% as.data.frame() %>%
  tibble::rownames_to_column(var = "viewing") %>%
  tidyr::pivot_longer(cols = c("16-34", "35-54", "54+"),
                      names_to = "age", values_to = "count")
p_base = ggplot(y, aes(x = age, y = count, fill = viewing)) +
  theme_bw(base_size = 12) +
  scale_fill_brewer(palette = "Set1") +
  labs(fill = "", x = "Age group") +
  theme(legend.position = "top")
p1 = p_base +
  geom_bar(stat = "identity") +
  labs(y = "Count")
p2 = p_base +
  geom_bar(stat = "identity", position = "fill") +
  labs(y = "Proportion")
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



1.4 Income and IQ

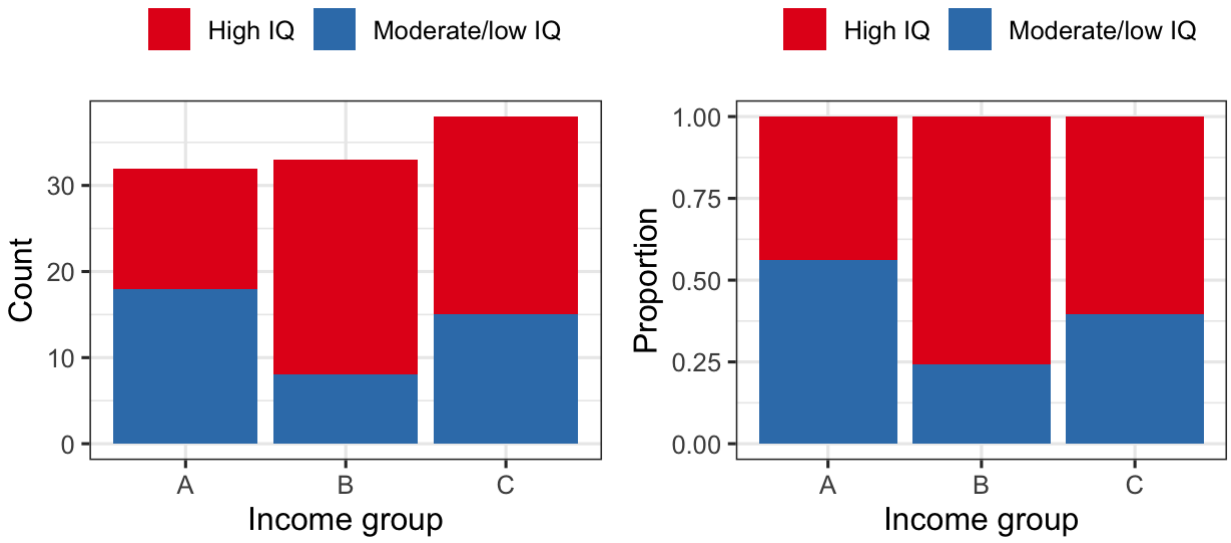
103 children attending a pre-school were classified by parents' income group and by IQ (intelligence quotient).

		High IQ	Moderate/low IQ
Income group	A	14	18
	B	25	8
	C	23	15

Does it look like the fractions of IQ differ significantly in the three income groups? No need to do a test

Does it look like the fractions of IQ differ significantly in the three income groups? No need to do a test at this point, just consider the observed counts, and the visualisations below.

```
library("tidyverse")
x = matrix(c(14, 25, 23, 18, 8, 15), ncol = 2)
colnames(x) = c("High IQ", "Moderate/low IQ")
rownames(x) = c("A", "B", "C")
y = x %>% as.data.frame() %>%
  tibble::rownames_to_column(var = "income") %>%
  tidyr::pivot_longer(c("High IQ", "Moderate/low IQ"),
                      names_to = "iq", values_to = "count")
p_base = ggplot(y, aes(x = income, y = count, fill = iq)) +
  theme_bw(base_size = 12) +
  scale_fill_brewer(palette = "Set1") +
  labs(fill = "", x = "Income group") +
  theme(legend.position = "top")
p1 = p_base +
  geom_bar(stat = "identity") +
  labs(y = "Count")
p2 = p_base +
  geom_bar(stat = "identity", position = "fill") +
  labs(y = "Proportion")
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



2 Group work

Discuss with your group:

- What does independence mean (in a statistical context)?
- Think of two things that are independent, explain why they are independent.
- How do you know they are independent?
- How does independence differ from homogeneity?

3 Exercises

3.1 Personality type

A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. She performs a personality test on a random sample of 258 business students and a random sample of 355 social science students. The results of the study are shown in the table below. What is the appropriate test in this context? [I.e. a test of goodness of fit, homogeneity or independence.] Perform the test using a 5% level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

```
counts = c(41, 52, 46, 61, 58, 72, 75, 63, 80, 65)
c_mat = matrix(counts, nrow = 2, byrow = TRUE)
colnames(c_mat) = c("Open", "Conscientious", "Extrovert", "Agreeable",
    "Neurotic")
rownames(c_mat) = c("Business", "Social Science")
```

3.2 Shocking

A psychological experiment was done to investigate the effect of anxiety on a person's desire to be alone or in company.

A group of 30 subjects was randomly divided into two groups of sizes 13 and 17.

The subjects were all told that they would be subject to electric shocks.

- The "high anxiety" group was told that the shocks would be quite painful
- The "low anxiety" group was told that they would be mild and painless

Both groups were told that there would be a 10 minute wait before the experiment began and each subject was given the choice of waiting alone or with other subjects.

The results were as follows:

	Wait together	Wait alone	Total
High anxiety	12	5	17
Low anxiety	4	9	13
Total	16	14	30

If we're picking between homogeneity and independence, which is more appropriate here?

At the 5% level of significance perform each of the following tests:

i. Fisher's exact test

ii. A chi-squared test without a continuity correction

iii. A chi-squared test with a continuity correction.

iv. A chi-squared test using a Monte Carlo p-value (i.e. using simulation).

Do the results of the different tests agree? Which are you most convinced by?

Would it make sense to calculate a relative risk here? Calculate the odds ratio, confidence interval and provide an interpretation.

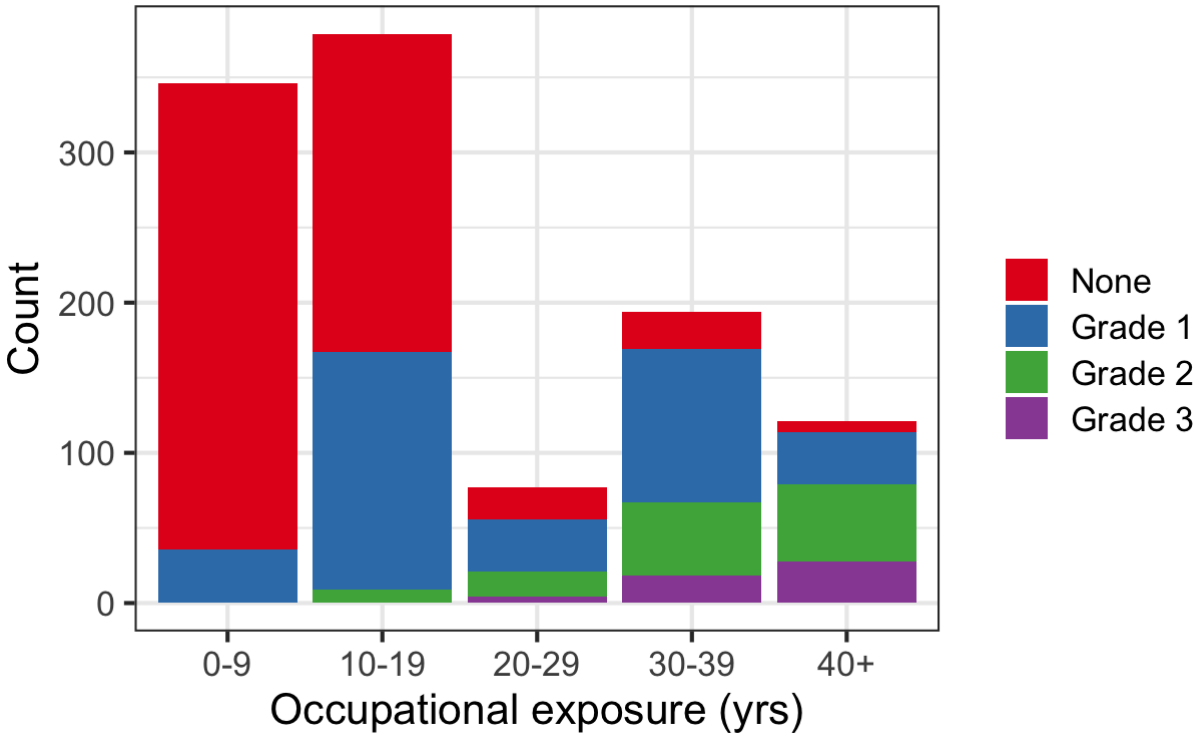
3.3 Asbestos

One of the breakthroughs that demonstrated the dangers to the exposure of asbestos is due to a study undertaken in the 1960's (data reported in Selikoff (1981)). Chest x-rays of a sample of 1117 workers in New York were taken to determine the damage done due to the occupational exposure of the workers to asbestos fibres. These workers were classified according to their years of exposure to the fibres and the severity of asbestosis that they were diagnosed with. The data appear in the following contingency table

Occupational exposure (yrs)	**Asbestos grade diagnosed**				
	None	Grade 1	Grade 2	Grade 3	Total
0-9	310	36	0	0	346
10-19	212	158	9	0	379
20-29	21	35	17	4	77
30-39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

```
asbestos = matrix(c(310, 212, 21, 25, 7, 36, 158, 35, 102, 35, 0, 9, 17, 49, 51, 0, 0, 4,
        18, 28), nrow = 5)
colnames(asbestos) = c("None", "Grade 1", "Grade 2", "Grade 3")
rownames(asbestos) = c("0-9", "10-19", "20-29", "30-39", "40+")
y = asbestos %>% as.data.frame() %>%
  tibble::rownames_to_column(var = "years") %>%
  tidyr::gather(key = grade, value = count, -years)
y$grade = factor(y$grade, levels = c("None", "Grade 1", "Grade 2", "Grade 3"), ordered =
        TRUE)
```

```r
ggplot(y, aes(x = years, y = count, fill = grade)) +
  geom_bar(stat = "identity") +
  theme_bw(base_size = 16) +
  scale_fill_brewer(palette = "Set1") +
  labs(fill = "", y = "Count", x = "Occupational exposure (yrs)")
```



1. Adapt the **ggplot2** code above such that the y-axis is a proportion within each exposure length group. Does it look like there's a relationship between the two variables?

2. Use the function `chisq.test()` to perform a standard chi-squared test of independence to determine whether there exists a statistically significant association between years of exposure to asbestos fibres and the severity of asbestosis that they were diagnosed with.

3. Use `x = r2dtable(____)` to randomly generate a contingency table with the same row and column totals as `asbestos`. Perform a chi-squared test and extract the test statistic using `chisq.test(x[[1]])$statistic`.

4. By using the `r2dtable()` function, perform a Monte-Carlo simulation to determine the p-value for the chi-squared test of independence. Generate 10,000 bootstrap resamples. Note: if doing this in an Rmd script, you might want to wrap your `chisq.test(___)$statistic` in `suppressWarnings()` so they don't slow down your computer, e.g. `suppressWarnings(chisq.test(___)$statistic)`. Plot a histogram of your Monte Carlo test statistics.

5. Use the `chisq.test()` function to perform a Monte-Carlo simulation that obtains a p-value. Do so using 10,000 bootstrap resamples.

4 Project

The first report will look at data from the class survey. The column names for the class survey are below. Break up into groups and discuss the following:

1. Which of these are categorical variables?
2. Which pairs of variables (if any) do you think would be related? [Doesn't have to be just categorical, could also start to think about whether there's a relationship between numeric variables or between a categorical variable and a numeric variable.]
3. What are some visual ways that you could communicate or explore these relationships?
4. What are some of the issues with the way the survey was written and how responses were recorded? Tabulate some of the categorical variables to see how these issues manifest. [Note that there's some overlap between this and your assignment, it's OK to discuss this with your group, so long as your submission is written in your own words.] We will talk about cleaning the data in one of the live lectures.
5. (Extension) In your own time you might want to perform hypothesis tests to check whether there's a statistically significant relationship (this is something you will need to do in your assignment).

The report is an individual assignment, but you're encouraged to discuss your approach with other students and your tutor. This is particularly valuable in an online setting - studying online there's not a lot of opportunity to talk through your thinking with other students. The tutor(s) will move from breakout room to breakout room and help clarify your thinking.

```
url = "https://docs.google.com/spreadsheets/d/1-
        DmA1UUM6QmZyucYiutuZX4Q0omtSCDwSOCNzHibkto/export?format=csv"
survey = readr::read_csv(url)
colnames(survey) %>%
    tibble() %>%
    gt::gt()
```

.
Timestamp
In the past 2 months, how many times have you had a COVID test?
What are your current living arrangements?
How tall are you?
If there is an event on Wednesday, and you are notified it has been moved forward 2 days, which day is the event?
Are you currently in Australia?

How do you self assess your mathematical ability?
How do you self assess your R coding ability?
How are you finding DATA2002 so far?
What year of university are you in?
How often do you turn your camera on in Zoom tutorials?
What's your COVID vaccination status?
What is your favourite social media platform?
Gender
How do you like your steak cooked?
What is your dominant hand?
On a scale from 0 to 10, please indicate how stressed you have felt in the past week.
On a scale from 0 to 10, please rate your current feeling of loneliness
How many non-spam emails did you receive to your University email account last Friday?
What do you typically say before signing off your name in an email?

What do you believe is the average entry salary in Australian Dollars of a data scientist who has just completed their undergraduate degree in data science?
Which unit are you enrolled in?
For which of your major(s) is this unit core or selective?
How many hours each week do you spend exercising?

5 For after the lab

5.1 IQ and Income

103 children attending a pre-school were classified by parents' income group and by IQ (intelligence

quotient).

Income group	High IQ	Moderate/low IQ
A	14	18
B	25	8
C	23	15

Do these data suggest that there is an association between income group and student IQ?

5.2 Eating habits and living arrangements

Consider the table below. It suggests that people that people who live with others are marginally more likely to be on a diet but are much less likely to watch what they eat and drink and are much more likely to eat and drink whatever they feel like. However, only 32 in the table are classified as living alone, so it is likely that these results reflect a relatively high degree of sampling error.

	Living alone	Living with others
On a diet	2 (6%)	25 (8%)
Watch what I eat and drink	23 (72%)	146 (49%)
Eat and drink whatever I feel like	7 (22%)	124 (42%)
Total	32 (100%)	295 (100%)

Perform a chi-squared test of homogeneity to see whether the apparent differences in a table like this are consistent with sampling error.

5.3 TV violence

A study of the amount of violence viewed on television as it relates to the age of the viewer yields the results shown in the accompanying table for a random sample of 81 people.

	Age		
Viewing	16 – 34	35 – 54	55 and over
Low violence	8	12	21
High violence	18	15	7

Do the data indicate that the viewing of violence is independent of age of viewer?

References

Selikoff, I. J. 1981. "Household Risks with Inorganic Fibers." *Bulletin of the New York Academy of Medicine* 57 (10): 947–61. https://doi.org/10.1177/1098214011426594.