Lab 01C: Week 4 (Solutions)

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

b.

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}

c.

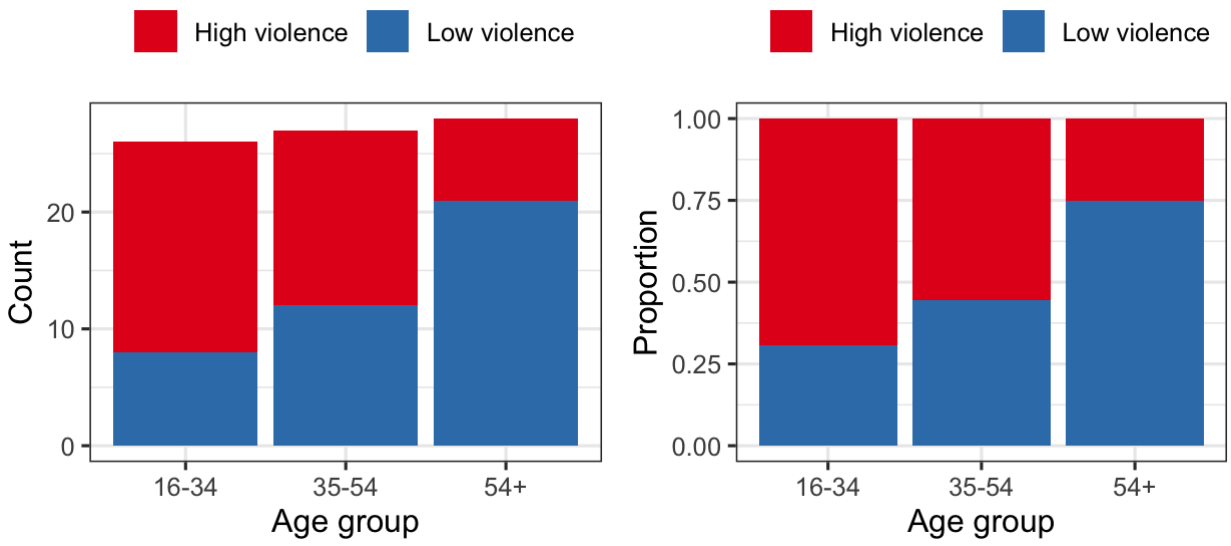
1.3 TV violence

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")
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

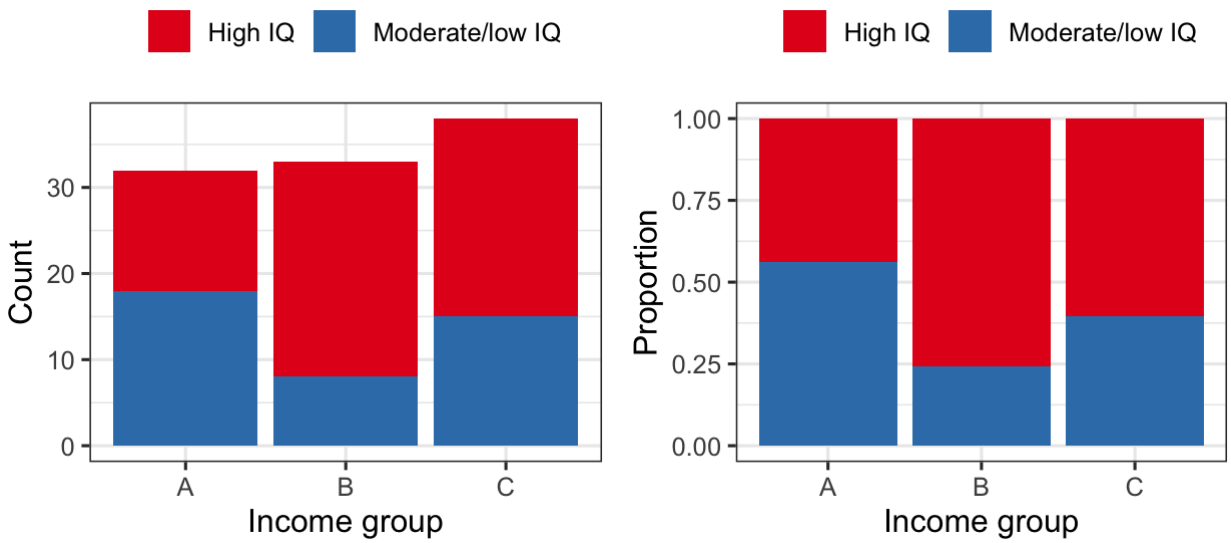
103 children attending a pre-school were classified by parents' income group and by IQ (intelligence quotient).

	High IQ	Moderate/low IQ
A	14	18

Income group	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 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")
```

```
chisq.test(c_mat)
```

```
    Pearson's Chi-squared test

data:  c_mat
X-squared = 3.006, df = 4, p-value = 0.5568
```

Check the expected cell counts:

```
chisq.test(c_mat)$expected %>%
    round(1)
```

```
               Open Conscientious Extrovert Agreeable Neurotic
Business       47.6          53.5      45.9      59.3     51.8
Social Science 65.4          73.5      63.1      81.7     71.2
```

H_0: The distribution of personality types is the same for both majors

H_1: The distribution of personality types is not the same for both majors

Assumptions: $e_{ij} \geq 5$ (confirmed by calculating the expected cell counts) and independent observations (confirmed as we are told there was random sampling from each population).

Test statistic: $T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approximately.

Observed test statistic: $t_0 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} = 3.006$

P-value: $P(T \geq t_0) = P(\chi^2_{(r-1)(c-1)} \geq t_0) = P(\chi^2_4 \geq 3.006) = 0.5568$.

```
1 - pchisq(3.006, 4)
```

```
[1] 0.5568218
```

Decision: Since the p-value much greater than 0.05, we do not reject H_0. There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors. Another way of saying this is: the data are consistent with the null hypothesis that the distribution of personality types is the same across business and social science majors.

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.

In this example, we started with one population, but then we stratified by anxiety, so in a sense we have two groups (sub-populations), one where we told them the shock would be quite painful and the other where we told them it would be mild. In this context it is more like a test for homogeneity where the null hypothesis is that the proportion of people who choose to wait alone is the same in both groups and the proportion of people who choose to wait together is the same in both groups.

```
counts = c(12, 5, 4, 9)
c_mat = matrix(counts, ncol = 2, byrow = TRUE)
colnames(c_mat) = c("Together", "Alone")
rownames(c_mat) = c("High", "Low")
c_mat
```

```
     Together Alone
High       12     5
Low         4     9
```

```
chisq.test(c_mat, correct = TRUE)
```

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

data:  c_mat
X-squared = 3.2294, df = 1, p-value = 0.07233
```

```
chisq.test(c_mat, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  c_mat
X-squared = 4.693, df = 1, p-value = 0.03029
```

```
fisher.test(c_mat)
```

```
    Fisher's Exact Test for Count Data
```

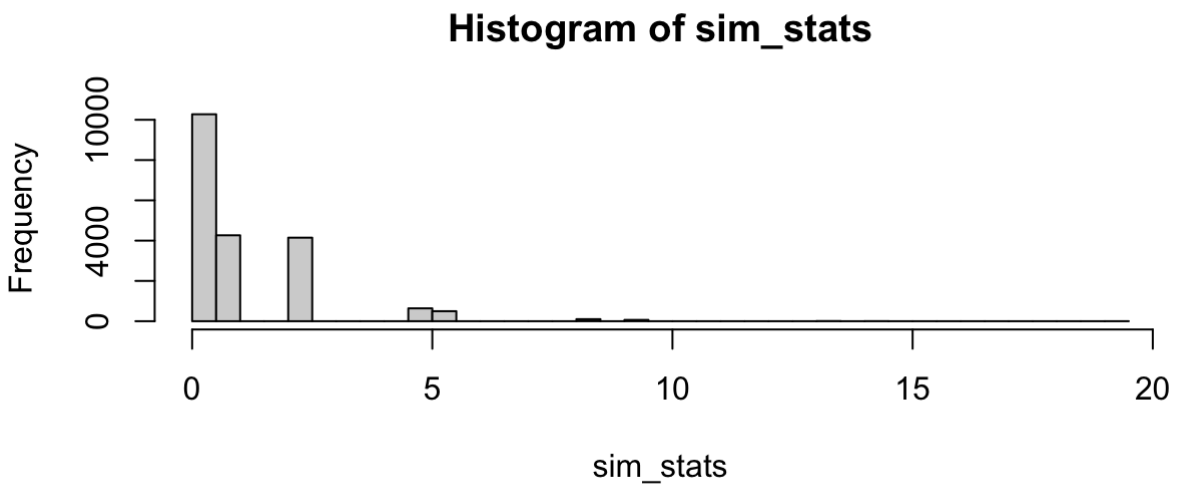
```
data:  c_mat
p-value = 0.06336
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.8958353 35.0773293
sample estimates:
odds ratio
  5.069118
```

```r
set.seed(1)
chisq.test(c_mat, simulate.p.value = TRUE, B = 20000)
```

```
    Pearson's Chi-squared test with simulated p-value (based on
    20000 replicates)

data:  c_mat
X-squared = 4.693, df = NA, p-value = 0.06425
```

```r
# if you want to do the simulation 'manually': 1. extract the test
# statistic from the original data
test_stat = chisq.test(c_mat, correct = FALSE)$statistic
# 2. generate 20000 tables with the same margins as the observed data
set.seed(2002)
rand_tables = r2dtable(n = 20000, r = rowSums(c_mat), c = colSums(c_mat))
# 3. calculate the the test statistic (without a continuity
# correction) for each of the randomly generated tables.  Notes:
# lapply() applys a function to each of the elements in a list
# unlist() takes a list and converts it to a vector
sim_stats = unlist(lapply(rand_tables, function(x) chisq.test(x, correct =
        FALSE)$statistic))
# 4. have a look at the distribution of the test statistics that were
# generated under the null hypothesis of independence
hist(sim_stats, breaks = 30)
```



```
# 5. calculate the Monte Carlo p-value as the proportion of simulated
# test statistics that are more extreme than the text statistic that
# we observed.
mean(sim_stats >= test_stat)
```

```
[1] 0.0657
```

Only the chi-squared test without the continuity correction gave a p-value that was less than 0.05. We're more convinced by the other approaches which give more reliable results, particularly when the sample sizes are small. The Monte Carlo p-value is very similar to Fisher's exact test (these would be our most preferred solutions) while the p-value for a chi-squared test with continuity correction is slightly larger.

For the odds ratio,

```
c_mat
```

```
     Together Alone
High       12     5
Low         4     9
```

```
mosaic::oddsRatio(c_mat, verbose = TRUE)
```

```
Odds Ratio

Proportions
       Prop. 1:  0.7059
       Prop. 2:  0.3077
     Rel. Risk:  0.4359

Odds
        Odds 1:  2.4
        Odds 2:  0.4444
    Odds Ratio:  0.1852

95 percent confidence interval:
     0.1824 < RR < 1.042
     0.0384 < OR < 0.8932
NULL
```

```
[1] 0.1851852
```

In this example we have sampled from the two groups (i.e. we fixed the number in the high group and we fixed the number in the alone group), so it makes sense to estimate the conditional probabilities $P(\text{Together} \mid \text{High})$ and $P(\text{Together} \mid \text{Alone})$.

If we interpret this output

If we interpret this output,

- `Prop. 1: 0.7059` is our estimate of $P(\text{Together} \mid \text{High})$, the proportion of subjects who preferred to wait together in the high anxiety group (12/(12+5)).
- `Prop. 2: 0.3077` is our estimate of $P(\text{Together} \mid \text{Low})$, is the proportion of subjects who preferred to wait together in the low anxiety group (4/(4+9)).
- The relative risk reported by this function is the ratio of these two conditional probabilities, `Rel. Risk: 0.4359` = `(Prop. 2)/(Prop. 1)` = `0.3077 / 0.7059` = 0.44. This is different to what we would have done from the lecture, where we would have calculated `(Prop. 1)/(Prop. 2)` = `0.7059 / 0.3077` = 2.3. Either way is OK so long as we adjust the interpretation.
  - For the 2.3 relative risk, we're saying that subjects who were told that it would be a painful shock were 2.3 times more likely to wait together than subjects who were told it wouldn't be painful.
  - For the 0.44 relative risk, we're saying that subjects who were told that it would not be a painful shock were 0.44 times more likely (i.e. they were less likely) to wait together than subjects who were told it would be painful.
- `Odds 1: 2.4` is our estimate of $O(\text{Together} \mid \text{High}) = P(\text{Together} \mid \text{High})/P(\text{Alone} \mid \text{High})$, the odds of subjects who preferred to wait together in the high anxiety group to (0.7059/(1-0.7059)).
- `Odds 2: 0.4444` is our estimate of $O(\text{Together} \mid \text{Low}) = P(\text{Together} \mid \text{Low})/P(\text{Alone} \mid \text{Low})$, the odds of subjects who preferred to wait together in the low anxiety group to (0.3077/(1-0.3077)).
- The odds ratio reported by the function is `Odds Ratio: 0.1852` = `Odds 2`/`Odds 1` = `0.4444/2.4`. If we were following the approach in the lecture slides we would have calculated `Odds 1`/`Odds 2` = `2.4/0.4444` = 5.4. Either way we just need to adjust out interpretation.
  - For the 5.4 odds ratio, we're saying that the odds of waiting together for the painful shock group are 5.4 times the odds of waiting together for the mild shock group.
  - For the 0.19 odds ratio, we're saying that the odds of waiting together for the mild shock group are 0.19 times the odds of waiting together for the painful shock group.
- The null hypothesis is that the odds ratio is equal to 1 (no association). The 95% confidence interval for the odds ratio, (0.0384, 0.8932) does not contain 1, therefore we would reject the null hypothesis. HOWEVER, remember that the calculation of the confidence interval for the odds ratio, relied on similar assumptions to the chi-squared test, i.e. we need "reasonably large" sample sizes in each of the cells (can think of this as the expected cell counts of at least 5 assumption).

Note: to get the same values as we would have calculated in lectures, we just need to flip the rows in the table:

```
mosaic::oddsRatio(c_mat[2:1, ], verbose = TRUE)
```

```
Odds Ratio

Proportions
       Prop. 1:  0.3077
       Prop. 2:  0.7059
     Rel. Risk:  2.294

Odds
        Odds 1:  0.4444
        Odds 2:  2.4
    Odds Ratio:  5.4

95 percent confidence interval:
     0.96 < RR < 5.483
     1.12 < OR < 26.04
NULL

[1] 5.4
```

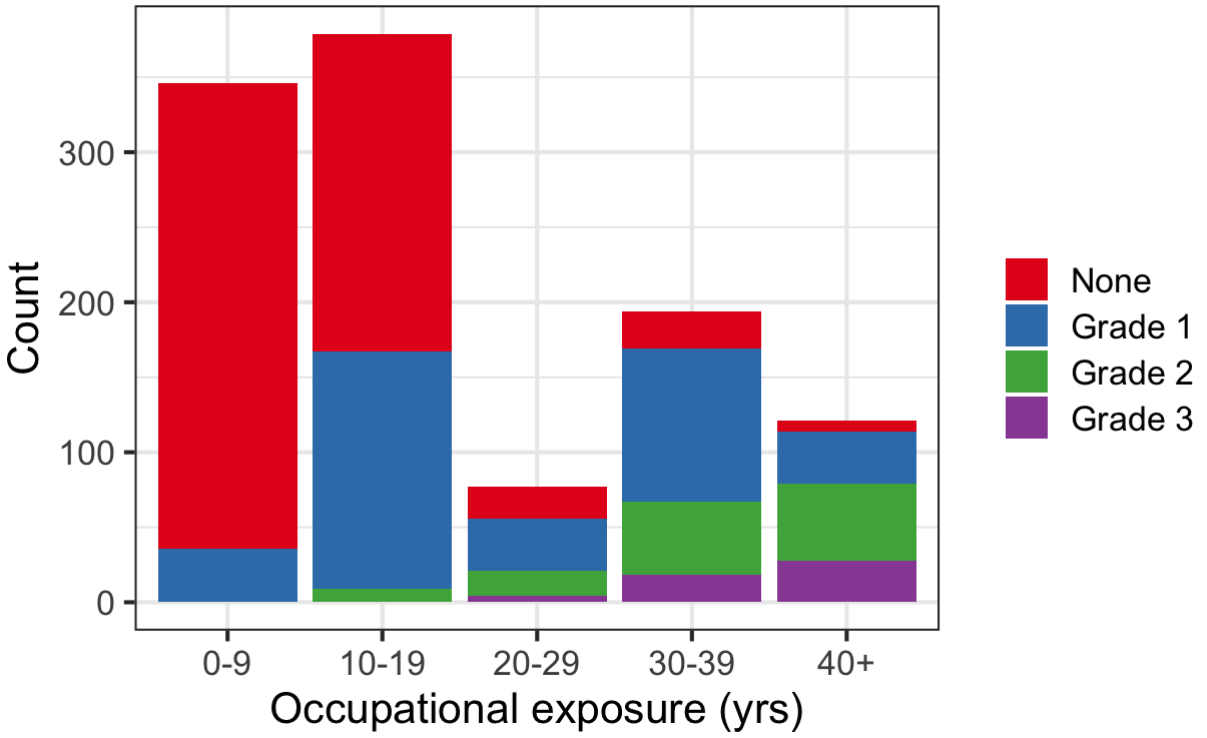
3.3 Asbestos

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)
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`.

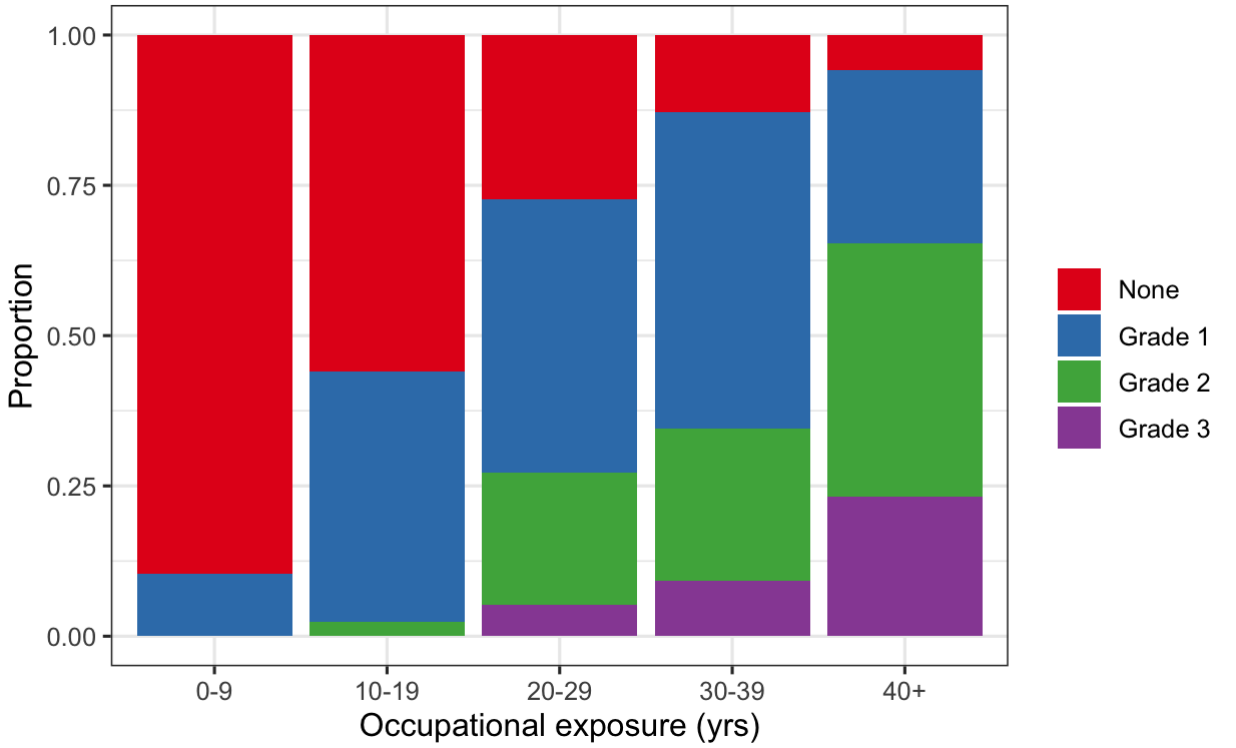
4. By using the `r2dtable()` function, perform a Monte-Carlo simulation to determine the p-value for the chi-squared test of independence. Generate 10,000 bootstrap resamples. Note: if doing this in an Rmd script, you might want to wrap your `chisq.test(___)$statistic` in `suppressWarnings()` so they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

1. Looks like there is a relationship between occupational exposure and asbestos grade. Longer exposure leads to higher grade.

```
ggplot(y, aes(x = years, y = count, fill = grade)) + geom_bar(stat = "identity",
    position = "fill") + theme_bw(base_size = 12) + scale_fill_brewer(palette = "Set1") +
    labs(fill = "", y = "Proportion", x = "Occupational exposure (yrs)")
```



2. The chi-squared test returns a very small p-value. Hence, there is evidence to suggest that a statistically significant association exists between exposure to asbestos fibres and the severity of asbestosis that a worker is diagnosed with.

```
chisq.test(asbestos)
```

```
Warning in chisq.test(asbestos): Chi-squared approximation may be
incorrect


    Pearson's Chi-squared test

data:  asbestos
X-squared = 648.81, df = 12, p-value < 2.2e-16
```

BUT there is a warning message indicating that the expected cell count assumption may not be met and so using the chi-squared distribution to compare the test statistic to may not be valid.

We can extract the expected cell counts as follows:

```
chisq.test(asbestos)$expected %>%
    round(1)
```

```
       None Grade 1 Grade 2 Grade 3
0-9   178.1   113.4    39.0    15.5
10-19 195.1   124.2    42.8    17.0
20-29  39.6    25.2     8.7     3.4
30-39  99.9    63.6    21.9     8.7
40+    62.3    39.6    13.6     5.4
```

There's just one cell that has a small expected cell count. Some text books would say this is OK [1], and it is probably not a huge issue in this case when most of the other expected cell counts are reasonable large.

However, an alternative approach is to perform a permutation test, where we still use the test statistic but we no longer compare it to a chi-squared distribution, rather we resample the data in such a way that we know the rows and columns are independent and assuming the marginal totals of the contingency table are fixed.

```
t0 = chisq.test(asbestos)$statistic
```

3. We start by calculating the row and column totals:

```
row_totals = rowSums(asbestos)
row_totals
```

```
 0-9 10-19 20-29 30-39   40+
 346   379    77   194   121
```

```
col_totals = colSums(asbestos)
col_totals
```

```
  None Grade 1 Grade 2 Grade 3
   575     366     126      50
```

Now we can use the `r2dtable()` function to randomly generate a contingency table with the same row and column totals:

```
set.seed(2018)
rnd = r2dtable(n = 1, r = row_totals, c = col_totals)
chisq.test(rnd[[1]])$statistic
```

```
X-squared
 7.150517
```

4. The Monte-Carlo p-value obtained by generating 10,000 contingency tables, computing the chi-squared test statistic for each table and seeing the proportion of these exceed the observed test statistic.

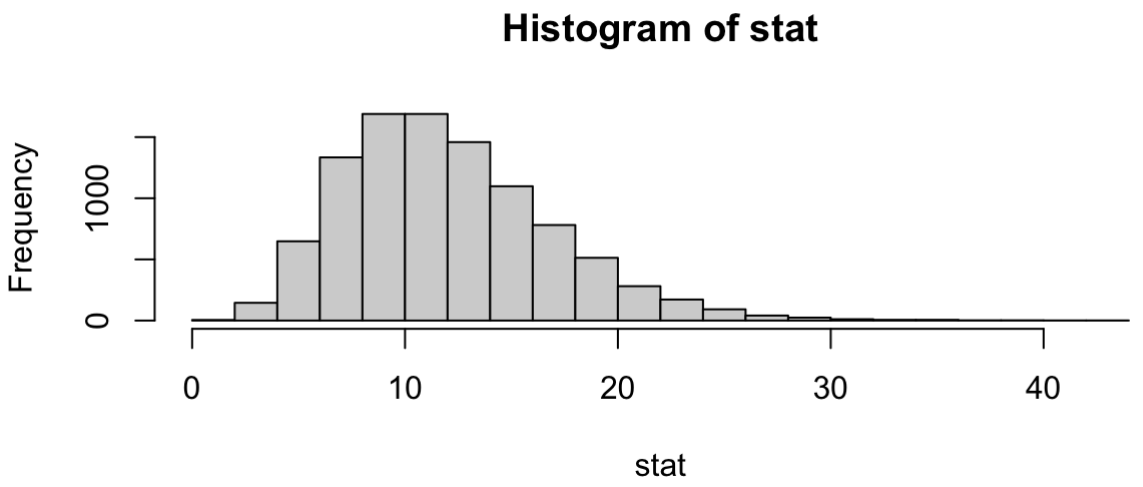
```
B = 10000
```

```
stat = numeric(length = B)
set.seed(2002)
tables = r2dtable(n = B, r = row_totals, c = col_totals)
for (i in 1:B) {
    stat[i] = suppressWarnings(chisq.test(tables[[i]], )$statistic)
}
stat = sapply(tables, function(x) suppressWarnings(chisq.test(x)$statistic))
mc_pval = mean(stat >= t0)
mc_pval
```

```
[1] 0
```

We can look at the distribution of test statistics:

```
hist(stat)
```



There are no permutation test statistics that were more extreme than the test statistic we observed on our original data. From the histogram, we can see that the observed test statistic, 648.8 is way past the range that we would expect to see if the null hypothesis of independence was true.

5. The `chisq.test()` function can do all this for us.

```
chisq.test(asbestos, simulate.p.value = TRUE, B = B)
```

```
    Pearson's Chi-squared test with simulated p-value (based on
    10000 replicates)

data:  asbestos
X-squared = 648.81, df = NA, p-value = 9.999e-05
```

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

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?

```
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")
chisq.test(x)
```

```
    Pearson's Chi-squared test

data:  x
X-squared = 6.9491, df = 2, p-value = 0.03098
```

Let $p_{1j}; j = 1, 2, 3$ denote the proportion of High IQ in the income group A, B, C respectively. Let $p_{2j}; j = 1, 2, 3$ denote the proportion of Moderate/low IQ in the income group A, B, C respectively.

The null hypotheses is the independence of IQ and income level and this can be written in symbols as $p_{ij} = p_{i\bullet}p_{\bullet j}$ for $i = 1, 2, 3$ and $j = 1, 2$. The alternative here is at least one of the equalities does not hold. The test statistics here is 6.95 with corresponding p-values < 0.05 using a χ^2-test with 2 degree for freedom. In summary, there is strong evidence in the data against H_0, i.e. there is IQ is not independent of income group.

Income		High IQ ($j=1$)	Low IQ ($j=2$)	Mar. prob. $p_{i\bullet}$
A		$y_{11}=14$	$y_{12}=18$	$y_{1\bullet}=32$
$(i=1)$	$e_{ij}=np_{i\bullet}p_{\bullet j}$	$103(0.602)(0.31)=19.26$	$103(0.398)(0.31)=12.74$	$\frac{32}{103}=0.311$
	$\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$	$\frac{(-5.26)^2}{19.26}=1.438$	$\frac{(5.26)^2}{12.7}=2.174$	
B		$y_{21}=25$	$y_{22}=8$	$y_{2\bullet}=33$
$(i=2)$	$e_{ij}=np_{i\bullet}p_{\bullet j}$	$103(0.602)(0.32)=19.86$	$103(0.398)(0.32)=13.14$	$\frac{33}{103}=0.320$
	$\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$	$\frac{(5.14)^2}{19.86}=1.328$	$\frac{(-5.14)^2}{13.1}=2.008$	
C		$y_{31}=23$	$y_{32}=15$	$y_{3\bullet}=38$
$(i=3)$	$e_{ij}=np_{i\bullet}p_{\bullet j}$	$103(0.602)(0.37)=22.87$	$103(0.398)(0.37)=15.13$	$\frac{38}{103}=0.369$
	$\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$	$\frac{(0.13)^2}{22.87}=0.001$	$\frac{(-0.13)^2}{15.1}=0.001$	
	$y_{\bullet j}=\sum_i y_{ij}$	62	41	103
	$\sum_i e_{ij}$	62.00	41.00	103
	$p_{\bullet j}$	$\frac{62}{103}=0.602$	$\frac{41}{103}=0.398$	

Hypotheses:

$$H_0:\ p_{ij}=p_{i\bullet}p_{\bullet j},\quad i=1,2,3,\ j=1,2\quad \text{vs}$$
$$H_1:\ \text{At least one of the equalities does not hold.}$$

or more simply, we can write H_0 : parents' income group and child's IQ group are independent vs H_1 : income group and IQ group are dependent.

Assumption: $e_{ij}=np_{i\bullet}p_{\bullet j}\geq 5$ and independent observations.

Test statistic: Under H_0, $T\sim \chi^2$ approx

Test statistic: Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approx.

Observed test statistic:

$$t_0 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(y_{ij} - y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}$$
$$= \frac{(14 - 19.26)^2}{19.26} + \ldots + \frac{(15 - 15.13)^2}{15.13} = 6.95$$

p-value: The corresponding p-value with $(r-1)(c-1) = 2$ degrees of freedom is $P(T \geq t_0) = P(\chi^2_2 \geq 6.949) = 0.031$.

Decision: Since the p-value less than 0.05, we reject H_0. There is evidence in the data to suggest that IQ is not independent of parents' income.

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.

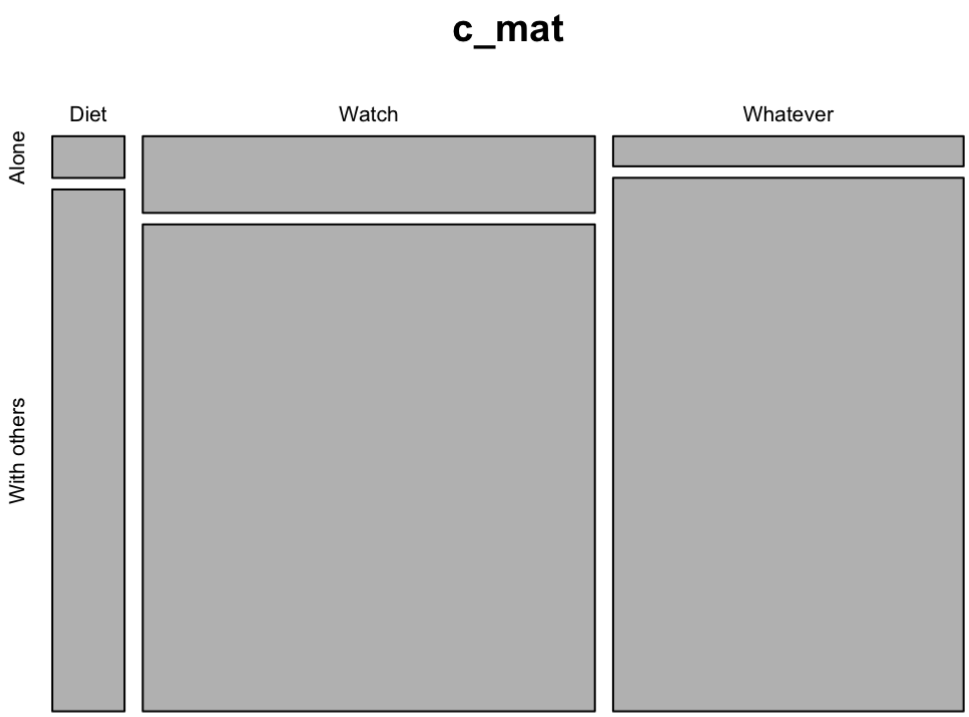
	Living alone	Living with others
On a diet	2 (6%)	25 (8%)
Watch what I eat and drink	23 (72%)	146 (49%)
Eat and drink whatever I feel like	7 (22%)	124 (42%)
Total	32 (100%)	295 (100%)

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

```
counts = c(2, 23, 7, 25, 146, 124)
c_mat = matrix(counts, ncol = 2, byrow = FALSE)
colnames(c_mat) = c("Alone", "With others")
rownames(c_mat) = c("Diet", "Watch", "Whatever")
c_mat
```

```
         Alone With others
Diet         2          25
Watch       23         146
Whatever     7         124
```

```
mosaicplot(c_mat)
```




```
chisq.test(c_mat)  # note low expected cell counts
```

```
    Pearson's Chi-squared test

data:  c_mat
X-squared = 5.9, df = 2, p-value = 0.05234
```

```
set.seed(2019)
chisq.test(c_mat, simulate.p.value = TRUE, B = 20000)
```

```
    Pearson's Chi-squared test with simulated p-value (based on
    20000 replicates)

data:  c_mat
X-squared = 5.9, df = NA, p-value = 0.0505
```

With this data, the p-value is pretty close to 0.05, so we don't have strong evidence one way or the other.

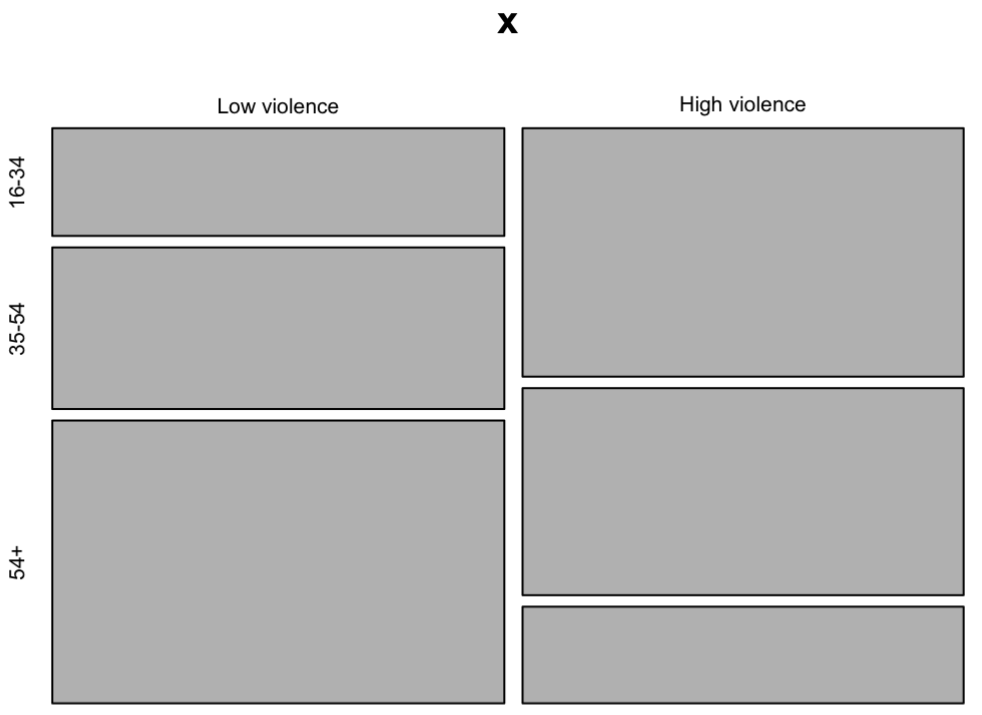
5.3 TV violence

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?

```
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")
mosaicplot(x)
```




```
chisq.test(x)
```

```
    Pearson's Chi-squared test

data:  x
```

```
X-squared = 11.169, df = 2, p-value = 0.003756
```

```
chisq.test(x)$expected
```

```
                16-34    35-54      54+
Low violence  13.16049 13.66667 14.17284
High violence 12.83951 13.33333 13.82716
```

Let $p_{1j}, j = 1, 2, 3,$ denote the probability of a person who is a Low violence viewer in the age group 16-34, 35-54, 55 and Over respectively,

and $p_{2j}, j = 1, 2, 3,$ denote the probability of a person who is a High violence viewer in the age group 16-34, 35-54, 55 and Over respectively.

The chi-squared test for independence between factors is as follows:

Hypotheses: H_0: viewing preference is independent of age group H_1: there is an association between viewing preference and age group.

Assumption: $e_i = np_{ij} \geq 5$ (verified by checking the expected cell counts) and independent observations (verified as we have a random sample).

Test statistic: $T = \sum_i \sum_j \frac{(Y_{ij} - e_{ij})}{e_{ij}}$ Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approximately.

Observed test statistic:

$$t_0 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(y_{ij} - y_{i\bullet} y_{\bullet j}/n)^2}{y_{i\bullet} y_{\bullet j}/n}$$
$$= \frac{(8 - 13.16)^2}{13.16} + \ldots + \frac{(7 - 13.83)^2}{13.83} = 11.169.$$

p-value: $P(T \geq t_0) = P(\chi^2_2 \geq 11.169) = 0.004.$

Decision: Since the p-value is less than 0.05, we reject H_0. There is strong evidence in the data that the view of violence is dependent on the age of the viewer.

Footnotes

1. Some text books say that the expected cell counts should be 5 or more in at least 80% of the cells, and no cell should

have an expected value of less than one [↩]

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.