Lab 01B: Week 3

Contents

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

- improve understanding of relative risk and odds ratios
- develop proficiency in performing chi-squared tests for goodness of fit

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 whether the proportion of births for DATA2002 students is 0.25 for each of the 4 seasons 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 confidence interval for the log odds ratio contains 1

1.2

In a test to see whether the proportion of births for DATA2002 students is 0.25 for each of the 4 seasons, assuming that the null hypothesis is true, the distribution of the test statistic is:

a. chi-squared with 3 degrees of freedom χ_3^2

b. chi-squared with 4 degrees of freedom χ_4^2

c. standard normal $Z \sim N(0, 1)$

d. t distribution with 3 degrees of freedom t_3

e. t distribution with 4 degrees of freedom t_4

1.3

A casino is worried about whether or not its die have been tampered with. To test this, a dealer rolls 4 dice 100 times and records the number of evens (2, 4 or 6) that appear.

Number of evens	0	1	2	3	4
Number of rolls of 4 dice	1	15	42	32	10

What distribution does the test statistic for a chi-squared goodness of fit follow in this example?

a. chi-squared with 1 degree of freedom χ_1^2

b. chi-squared with 2 degrees of freedom χ_2^2

c. chi-squared with 3 degrees of freedom χ_3^2

d. chi-squared with 4 degrees of freedom χ_4^2

e. chi-squared with 5 degrees of freedom χ_5^2

2 Group exercise

In week 2 we covered odds-ratios and relative risk. Within your group discuss:

- What are the key differences between prospective and retrospective study?
- What are relative risks? What are odds-ratios?
- Why would you use one over the other?

3 Exercises

3.1 Dishonest dice

A casino is worried about whether or not its die have been tampered with. To test this, a dealer rolls 4 dice 100 times and records how many even numbers (2, 4 or 6) appear.

Number of evens	0	1	2	3	4
Number of rolls of 4 dice	1	15	42	32	10

Can the scientist infer at the 5% significance level that the number of even when $n = 4$ dice are rolled follows a binomial random variable with $p = 1/2$? Recall, if $X \sim B(n, p)$ then $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$.

```
y = c(1, 15, 42, 32, 10)  # input the observed counts
x = 0:4  # define the corresponding groups
```

3.2 Mammograms

Suppose that among 100,000 women with negative mammograms, 20 will have breast cancer diagnosed within 2 years; and among 100 women with positive mammograms, 10 will have breast cancer diagnosed within 2 years. Clinicians would like to know if there is a relationship between a positive or negative mammogram and developing breast cancer?

Mammogram \ Breast cancer	Yes	No
Positive	10	90
Negative	20	99, 980

```
x = matrix(c(10, 20, 90, 99980), ncol = 2)
colnames(x) = c("Breast cancer: yes", "Breast cancer: no")
rownames(x) = c("Mammogram: positive", "Mammogram: negative")
```

1. Is it appropriate to use a relative risk to quantify the relationship between the risk factor (Mammogram result) and disease (Breast cancer)? If so calculate the relative risk and provide an interpretation.
2. Calculate the odds ratio of having breast cancer for positive vs negative mammograms and provide an interpretation.
3. Calculate a confidence interval for the odds-ratio, is there evidence that there might be a relationship between mammogram test results and breast cancer diagnosis?

3.3 Soccer goals

Goals per soccer game arrive at random moments, and could be reasonably modelled by a Poisson process. If so, the total number of goals scored in a soccer game should be a Poisson random variable.

Here are the number of goals scored in each of the $n = 104$ games at the 2015 FIFA Women's World Cup (source):

```
goals <- c(1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 2, 2, 4, 0, 10, 0, 1, 1, 2, 3,
    0, 4, 1, 3, 6, 0, 1, 0, 10, 1, 2, 1, 0, 1, 1, 2, 3, 3, 3, 1, 2, 0,
    0, 0, 0, 1, 1, 1, 1, 1, 2, 0, 1, 0, 2, 2, 0, 1, 2, 1, 1, 0, 1, 1, 0,
    2, 2, 1, 0, 5, 2, 1, 4, 1, 1, 0, 0, 1, 3, 0, 1, 0, 1, 2, 2, 0, 2, 1,
    1, 1, 0, 1, 0, 1, 2, 1, 2, 0, 2, 1, 0, 1, 5, 2)
observed_goals = table(goals)
```

Test the null hypothesis that the number of goals scored per game follows a Poisson distribution.

You will need to estimate the λ parameter and collapse categories (if necessary) to make sure the assumptions are met.

3.4 Education

This dataset measures the educational attainment of Americans by age categories in 1984. Counts are presented in thousands. Data collected by the U.S. Bureau of the Census. Americans under age 25 are not included because many have not completed their education. The variables are:

- `Education`: Level of education achieved
- `Age_Group`: Age group (years)
- `Count`: 1000's of Americans in this education and age category

Read in the data and check the size of your data. Think about what the number of rows actually means.

```r
## Reading in the data
library("tidyverse")
edu = readr::read_delim("https://raw.githubusercontent.com/DATA2002/data/master/education-by-age-census.txt",
    delim = "\t")
edu = edu %>%
    janitor::clean_names()
knitr::kable(edu)
```

We can summarise this data in a more “human friendly” format using the `tidyr::spread()` function:

```r
edu %>%
    tidyr::spread(key = age_group, value = count)
```

```
# A tibble: 4 × 6
  education                  `>64` `25-34` `35-44` `45-54` `55-64`
  <chr>                      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 College,1-3 years           2503    8555    5576    3124    2524
2 College,4 or more years     2483    9771    7596    3904    3109
3 Completed high school       7558   16431    1855    9435    8795
4 Did not complete high school 13746    5416    5030    5777    7606
```

```r
# an alternative approach is the xtabs function xtabs(count ~
# education + age_group, data = edu)
```

Note that the categories aren’t in a sensible order, let’s reorder (relevel) them. To do this we’ll use the **forcats** package that is part of the **tidyverse**.

```r
edu = edu %>%
  dplyr::mutate(
    age_group = forcats::fct_relevel(age_group, ">64", after = 4),
    education = forcats::fct_relevel(education,
                                     "Did not complete high school",
                                     "Completed high school",
                                     "College,1-3 years",
                                     "College,4 or more years"))
tab = edu %>% tidyr::spread(key = age_group, value = count)
# tab = xtabs(count ~ education + age_group, data = edu)
tab
```

```
# A tibble: 4 × 6
  education                    `25-34` `35-44` `45-54` `55-64` `>64`
  <fct>                          <dbl>   <dbl>   <dbl>   <dbl> <dbl>
1 Did not complete high school    5416    5030    5777    7606 13746
2 Completed high school          16431    1855    9435    8795  7558
3 College,1-3 years               8555    5576    3124    2524  2503
4 College,4 or more years         9771    7596    3904    3109  2483
```

Many of the questions below are about college vs non-college. Let's add in a new variable in our data frame that identifies the college vs non-college categories.

```r
edu = edu %>%
    mutate(college = dplyr::if_else(stringr::str_detect(education, "College"),
        "College", "No college"))
```

And let's make a aggregated data frame `edu_college` that summarises over the different education levels, leaving totals for the `college` variable.

```r
edu_college = edu %>%
    dplyr::group_by(age_group, college) %>%
    dplyr::summarise(count = sum(count)) %>%
    dplyr::ungroup()
```

1. Which age category has the highest percentage of college graduates?
2. What percent of all Americans over age 25 never went to college?
3. Based on this data, is there evidence of a relationship between age category and educational attainment? In other words, is there evidence that younger people are more likely to have finished college than older people? Use graphical representation to compare the percent of people in each age group who have completed college. What is the appropriate statistical test to use here?

▶ **Hints**

4 For after the lab

See Larsen and Marx (2012) section 10.3 and 10.4 for further examples of goodness of fit tests when all parameters are known and when you need to estimate parameters.

4.1 Recap

R functions used:

- `sum()`, `length()`
- `dbinom()`, `pchisq()`, `qchisq()`, `qnorm()`
- `chisq.test()`
- `mosaic::oddsRatio()`
- `readr::read_delim()`
- `janitor::clean_names()`
- `tidyr::spread()`

- `tidyr::spread()`
- `dplyr::mutate()`
- `forcats::fct_relevel()`
- `dplyr::if_else()`
- `stringr::str_detect()`
- `dplyr::group_by()` and `dplyr::ungroup()`
- `ggplot()` with `geom_bar()`

4.2 Heart attacks and smoking

A group of 200 people who have experienced a heart attack and 200 with no heart attack were asked if they were ever smokers.

The results are presented in the table below:

Smoked \ Heart attack	Yes	No
Yes	33	18
No	167	182

1. Is it appropriate to use a relative risk to quantify the relationship between the risk factor (Smoking) and disease (Heart attack)? If so calculate the relative risk.
2. Calculate and interpret the odds ratio of having a heart attack for smokers compared to non-smokers.
3. Calculate a confidence interval for the odds ratio, is there evidence that there might be a relationship between smoking and heart attacks?

```r
x = matrix(c(33, 167, 18, 182), ncol = 2)
colnames(x) = c("Heart attack: yes", "Heart attack: no")
rownames(x) = c("Smoke: yes", "Smoke: no")
```

References

Larsen, Richard J., and Morris L. Marx. 2012. *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. Boston, MA: Prentice Hall.