Lab 01B: Week 3 (Solutions)

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

a.

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^2_3

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

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

a.

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^2_1

b. chi-squared with 2 degrees of freedom χ^2_2

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

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

e. chi-squared with 5 degrees of freedom χ^2_5

d. We didn't need to estimate any parameters, so our chi-squared test degrees of freedom is the number of categories minus 1. Also don't need to collapse any categories, because the expected cell counts will all be greather than 5, but this is shown later on. Can check it pretty easily by considering the probability of getting no evens (or all evens) as $(1/2)^4 = 1/16$ and then multiplying that by 100 to get the smallest expected cell count of $100/16 = 6.25$.

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
```

```
n = sum(y)  # number of rolls of 4 dice (sample size)
k = length(y)  # number of groups
p = dbinom(x, size = 4, prob = 1/2)  # obtain the p_i from the binomial pmf
p
```

```
[1] 0.0625 0.2500 0.3750 0.2500 0.0625
```

```
(ey = n * p)  # calculate the expected frequencies
```

```
[1]  6.25 25.00 37.50 25.00  6.25
```

```
ey >= 5  #check assumption e_i >= 5
```

```
[1] TRUE TRUE TRUE TRUE TRUE
```

```
(t0 = sum((y - ey)^2/ey))  # test statistic
```

```
[1] 13.16
```

```
(pval = 1 - pchisq(t0, df = k - 1))  # p-value
```

```
[1] 0.0105199
```

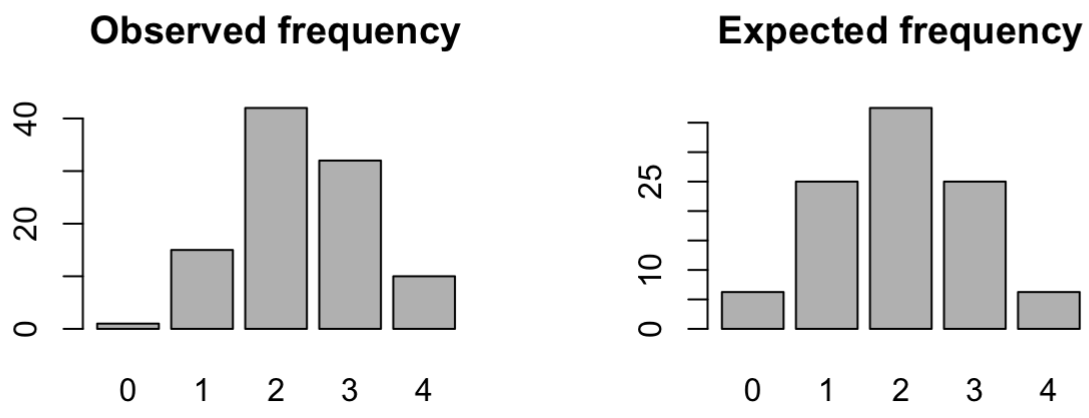
```
chisq.test(y, p = p)
```

```

    Chi-squared test for given probabilities

data:  y
X-squared = 13.16, df = 4, p-value = 0.01052
```

```
par(mfrow = c(1, 2))  # plot options
barplot(y, names.arg = x, main = "Observed frequency")
barplot(ey, names.arg = x, main = "Expected frequency")
```



The calculation is summarised in the following table:

No. x_i	Obs. freq. y_i	Exp. prob. $p_i = \binom{n}{x_i} p_0^{x_i}(1-p_0)^{5-x_i}$	Exp. freq. $e_i = np_i$	Chi-squared $\frac{(y_i-e_i)^2}{e_i}$
0	1	$\binom{4}{0} = 0.0625$	$100 \times 0.0625 = 6.25$	$\frac{(1-6.25)^2}{6.25} = 4.41$
1	15	$\binom{4}{1}0.5^1 0.5^3 = 0.2500$	$100 \times 0.2500 = 25.00$	$\frac{(15-25.00)^2}{25.00} = 4.00$
2	42	$\binom{4}{2}0.5^2 0.5^2 = 0.3750$	$100 \times 0.3750 = 37.50$	$\frac{(42-37.50)^2}{37.50} = 0.54$
3	32	$\binom{4}{3}0.5^3 0.5^1 = 0.2500$	$100 \times 0.2500 = 25.00$	$\frac{(32-25.00)^2}{25.00} = 1.96$
4	10	$\binom{4}{4}0.5^4 0.5^0 = 0.0625$	$100 \times 0.0625 = 6.25$	$\frac{(10-6.25)^2}{6.25} = 2.25$
Sum	100	1.0000	100.00	13.16

Let X be a random variable representing the number of boys in a family with 4 children. The chi-squared goodness-of-fit test to test if X follows a binomial distribution with $p = 1/2$ is

1. **Hypothesis:** H_0: X follows a binomial distribution with success probability $p = 1/2$ vs H_1: X does not follow a binomial distribution with $p = 1/2$.

2. **Assumptions:** independent observations (independent rolls of the 4 dice) and $e_i = np_i \geq 5$ (confirmed in the table above).

3. **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. Under H_0, $T \sim \chi^2_{k-1}$ approx.

4. **Observed test statistic:** $t_0 = 13.16$

5. **p-value:** $P(\chi^2_4 \geq 13.16) = 0.01052$

6. **Decision:** Since the p-value is less than 0.05, we reject the null hypothesis. The data is not consistent with the null hypothesis that the data follow a binomial distribution with probability of

consistent with the null hypothesis that the data follow a binomial distribution with probability of success, $p = 0.5$.

Aside: this is a sensible conclusion as the data were actually generated using:

```
set.seed(10)
y = table(rbinom(n = 100, size = 4, prob = 0.55))
```

In general, the closer the "alternative parameter" is to the hypothesised parameter, the larger the sample size you will need to be able to (correctly) reject the null hypothesis. Thought experiment: think about how hard it would be to reject the null hypothesis $H_0: if the data was generated using `prob = 0.51` vs how easy it would be to reject the null hypothesis if the data was generated using `prob = 0.7`.

Note, in an exam you might be given some R code such as this:

```
qchisq(c(0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99), 4) %>%
    round(3)
```

```
[1]  0.297  0.484  0.711  1.064  7.779  9.488 11.143 13.277
```

```
qchisq(c(0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99), 5) %>%
    round(3)
```

```
[1]  0.554  0.831  1.145  1.610  9.236 11.070 12.833 15.086
```

And be expected to be able to identify the relevant parts to make your conclusion.

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?

1. It is appropriate to use relative risk here as the study is prospective in nature. The participants were enrolled by risk factor (mammogram) and not the disease (breast cancer).

$$RR = \frac{a(c+d)}{c(a+b)} = \frac{10(20+99980)}{20(10+90)} = 500$$

This is very far from 1. Women with a positive mammogram are 500 times more likely to develop breast cancer than women with a negative mammogram.

```
# install.packages('mosaic')
1/mosaic::relrisk(x)
```

```
[1] 500
```

2. The odds ratio of developing breast cancer after a positive vs negative mammogram are

$$OR = \frac{ad}{cb} = \frac{10 \times 99980}{20 \times 90} = 555.4$$

```
1/mosaic::oddsRatio(x)
```

```
[1] 555.4444
```

We could interpret this as the odds of developing breast cancer after a positive mammogram are 555.4 times the odds after a negative mammogram.

Alternatively, we could say the odds of developing breast cancer is 555.4 higher given a positive mammogram compared to a negative mammogram result.

3. $$SE(\log(OR)) = \sqrt{1/10 + 1/90 + 1/20 + 1/99980} = 0.4$$

so the 95% confidence interval for log odds-ratio is $6.3 \pm 1.96 * 0.4 \approx (5.52, 7.08)$ and the confidence interval for the odds-ratio is there for $(e^{5.52}, e^{7.08}) = (248.6, 1192.73)$. Importantly, the value of 1 does not lie in this CI so we can conclude that there is a statistically significant association between the risk and the disease (at a 5% level of significance).

```
se = sqrt(1/10 + 1/90 + 1/20 + 1/99980)
or = (10 * 99980)/(20 * 90)
log_ci = c(log(or) - qnorm(0.975) * se, log(or) + qnorm(0.975) * se)
ci = exp(log_ci)
ci
```

```
CI
```

```
[1]  252.9119 1219.8657
```

Note that the difference between this and the interval calculated in the paragraph above are due to rounding errors.

Using the **mosaic** package:

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

```

Odds Ratio

Proportions
	   Prop. 1:	 0.1
	   Prop. 2:	 2e-04
	 Rel. Risk:	 0.002

Odds
	    Odds 1:	 0.1111
	    Odds 2:	 2e-04
	Odds Ratio:	 0.0018

95 percent confidence interval:
	 0.0009606 < RR < 0.004164
	 0.0008198 < OR < 0.003954
NULL

[1] 0.00180036
```

To get the same results as our manual calculations, we need to switch the rows:

```
y = x[c(2, 1), ]
mosaic::oddsRatio(y, verbose = TRUE)
```

```

Odds Ratio

Proportions
	   Prop. 1:	 2e-04
	   Prop. 2:	 0.1
	 Rel. Risk:	 500

Odds
	    Odds 1:	 2e-04
	    Odds 2:	 0.1111
	Odds Ratio:	 555.4

95 percent confidence interval:
	 240.2 < RR < 1041
	 252.9 < OR < 1220
NULL
```

```
[1] 555.4444
```

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.

We can fit a Poisson random variable with mean parameter λ calculated as:

```
(lambda = mean(goals))
```

```
[1] 1.403846
```

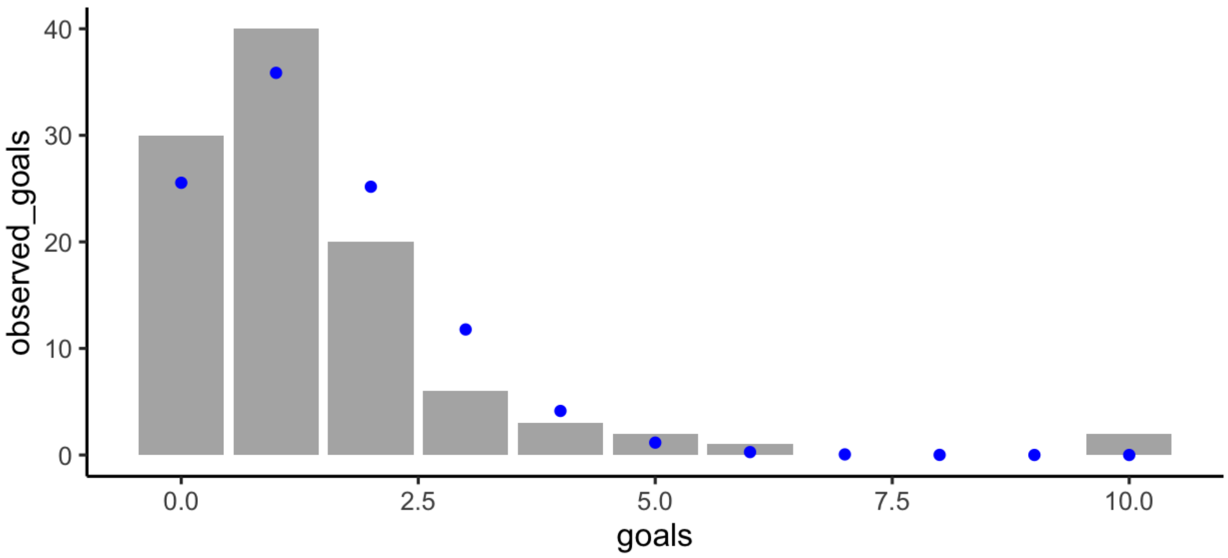
And work out the expected cell counts as follows:

```
hyp_probs = c(dpois(0:9, lambda), ppois(9, lambda, lower.tail = FALSE))
expected_goals = 104 * hyp_probs
round(expected_goals, 2)
```

```
 [1] 25.55 35.86 25.17 11.78  4.13  1.16  0.27  0.05  0.01  0.00  0.00
```

Let's take a look at the expected cell counts vs the actual cell counts. The bars are the observed counts and the dots are the expected cell counts under the null hypothesis of a Poisson distribution.

```
soccer_df = tibble(goals = 0:10, hyp_probs, expected_goals, observed_goals = c(30,
    40, 20, 6, 3, 2, 1, 0, 0, 0, 2))
soccer_df %>%
    ggplot() + aes(x = goals) + geom_col(aes(y = observed_goals), alpha = 0.5) +
    geom_point(aes(y = expected_goals), col = "blue")
```



We need to amalgamate the categories with small expected cell counts:

```
soccer_combo = soccer_df %>%
    slice(5:n()) %>%
    mutate(goals = "4+") %>%
    group_by(goals) %>%
    summarise(across(where(is.numeric), sum))
soccer_df2 = soccer_df %>%
    slice(1:4) %>%
    mutate(goals = as.character(goals)) %>%
    bind_rows(soccer_combo)
soccer_df2 %>%
    gt::gt() %>%
    gt::fmt_number(columns = 2, decimals = 3) %>%
    gt::fmt_number(columns = 3, decimals = 1)
```

goals	hyp_probs	expected_goals	observed_goals
0	0.246	25.5	30
1	0.345	35.9	40
2	0.242	25.2	20
3	0.113	11.8	6
4+	0.054	5.6	8

After amalgamating the categories, we're left with 5 goal outcomes (0, 1, 2, 3 and 4+) and we have estimated one parameter from the data (λ, the mean parameter of the Poisson random variable) so our test statistic will follow a chi-squared distribution with $5-1-1=3$ degrees of freedom, χ^2_3.

```
soccer_df2 = soccer_df2 %>%
    mutate(chi_sq = (observed_goals - expected_goals)^2/expected_goals)
```

```r
t0 = soccer_df2 %>%
    pull(chi_sq) %>%
    sum()
```

The observed test statistic is $t_0 = 6.148$.

```r
1 - pchisq(t0, df = 3)
```

```
[1] 0.1046487
```

The p-value is 0.1046 which is larger than 0.05, so we do not reject the null hypothesis at the 5% level of significance and conclude that the data are consistent with a Poisson distribution.

We could also use the `chisq.test()` function to calculate the test statistic:

```r
chisq.test(x = soccer_df2$observed_goals, p = soccer_df2$hyp_probs)
```

```

	Chi-squared test for given probabilities

data:  soccer_df2$observed_goals
X-squared = 6.1475, df = 4, p-value = 0.1884
```

The test statistic is correct, but the p-value is not right because it is based on an incorrect degrees of freedom - the `chisq.test()` function doesn't know that we've estimated a parameter from the data. Using this approach, we would need another step to calculate the p-value:

```r
results = chisq.test(x = soccer_df2$observed_goals, p = soccer_df2$hyp_probs)
t0 = results$statistic %>%
    unname()  # removes the name
t0
```

```
[1] 6.147538
```

```r
1 - pchisq(t0, df = 3)
```

```
[1] 0.1046487
```

```r
# or equivalently
pchisq(t0, df = 3, lower.tail = FALSE)
```

```
[1] 0.1046487
```

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.

```
## 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:

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

```
# A tibble: 4 x 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
```

```
# 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**.

```
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?

If we're practicing our **tidyverse** ninja skills,

```r
edu_college %>%
    group_by(age_group) %>%
    mutate(pct_in_age_grp = round(count/sum(count), 2) * 100) %>%
    arrange(college, age_group)
```

```
# A tibble: 10 × 4
# Groups:   age_group [5]
   age_group college    count pct_in_age_grp
   <fct>     <chr>      <dbl>          <dbl>
 1 25-34     College    18326             46
 2 35-44     College    13172             66
 3 45-54     College     7028             32
 4 55-64     College     5633             26
 5 >64       College     4986             19
 6 25-34     No college 21847             54
 7 35-44     No college  6885             34
 8 45-54     No college 15212             68
 9 55-64     No college 16401             74
10 >64       No college 21304             81
```

Alternatively using the tab object, we can identify the rows of interest, sum down those columns and divide those by the column totals:

```r
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
```

```r
colSums(tab[3:4, -1])/colSums(tab[, -1])
```

```
    25-34     35-44     45-54     55-64       >64 
0.4561770 0.6567283 0.3160072 0.2556504 0.1896539 
```

Ans: age group 35-44 with 66%

2. What percent of all Americans over age 25 never went to college?

```r
edu_college %>%
    group_by(college) %>%
    summarise(count = sum(count)) %>%
    mutate(pct = count/sum(count))
```

```
# A tibble: 2 × 3
  college    count   pct
  <chr>      <dbl> <dbl>
1 College    49145 0.376
2 No college 81649 0.624
```

```r
x = rowSums(tab[, -1])
sum(x[1:2]/sum(x))
```

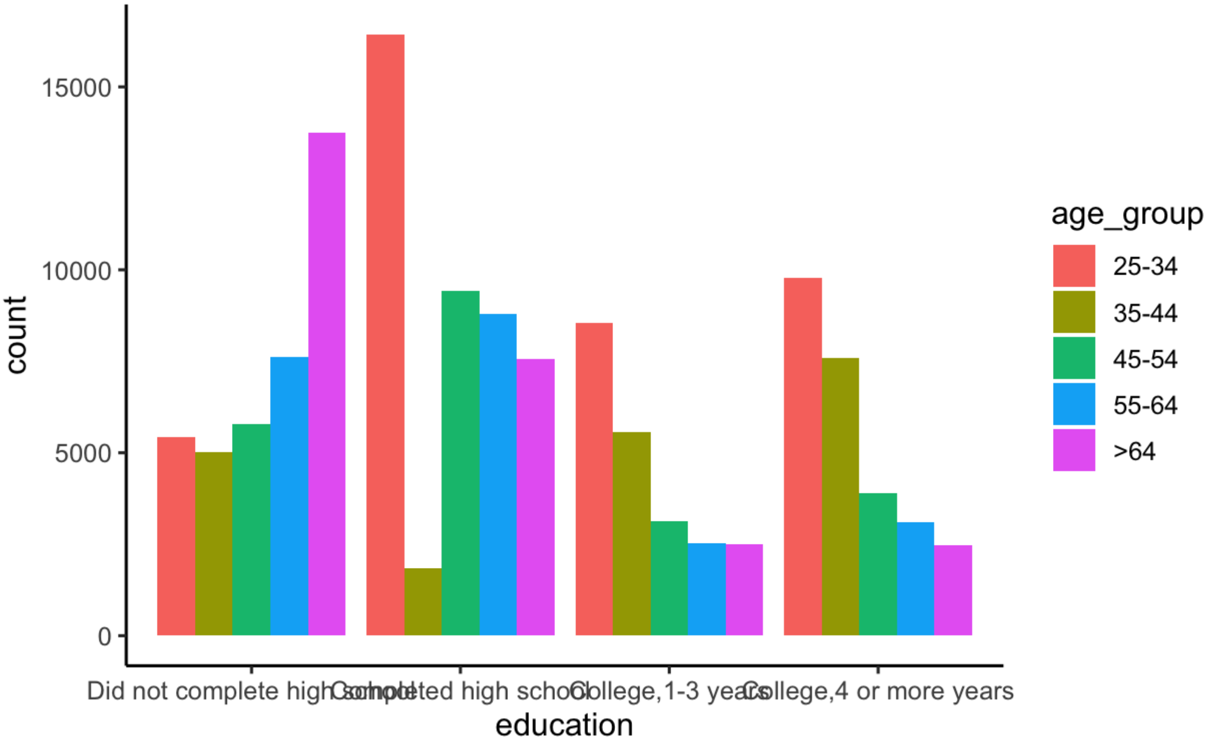
```
[1] 0.6242565
```

Answer: 62%

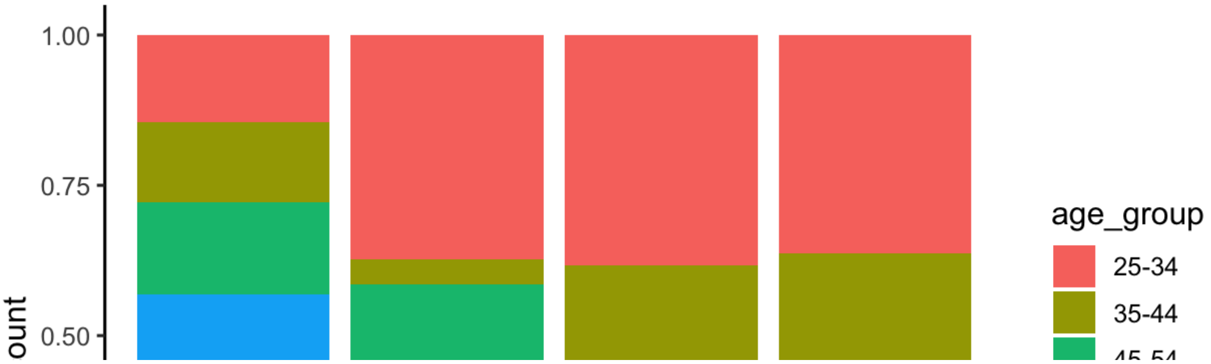
3. Based on this data, is there evidence of a relationship between age category and educational attainment? In other words, is there evidence that younger people are more likely to have finished college than older people? Use graphical representation to compare the percent of people in each age group who have completed college. What is the appropriate statistical test to use here?

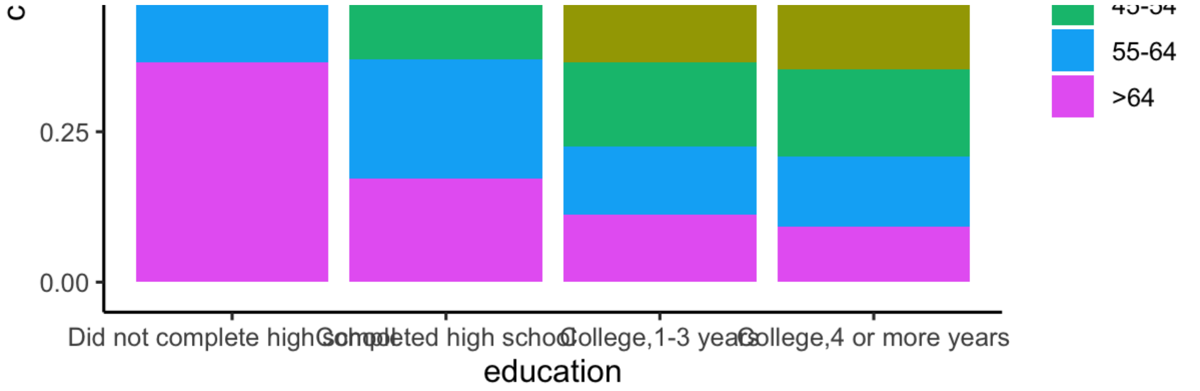
▶ **Hints**

```r
## Using education on the x-axis
ggplot(edu, aes(x = education, y = count, fill = age_group)) + geom_bar(stat = "identity",
    position = position_dodge())
```

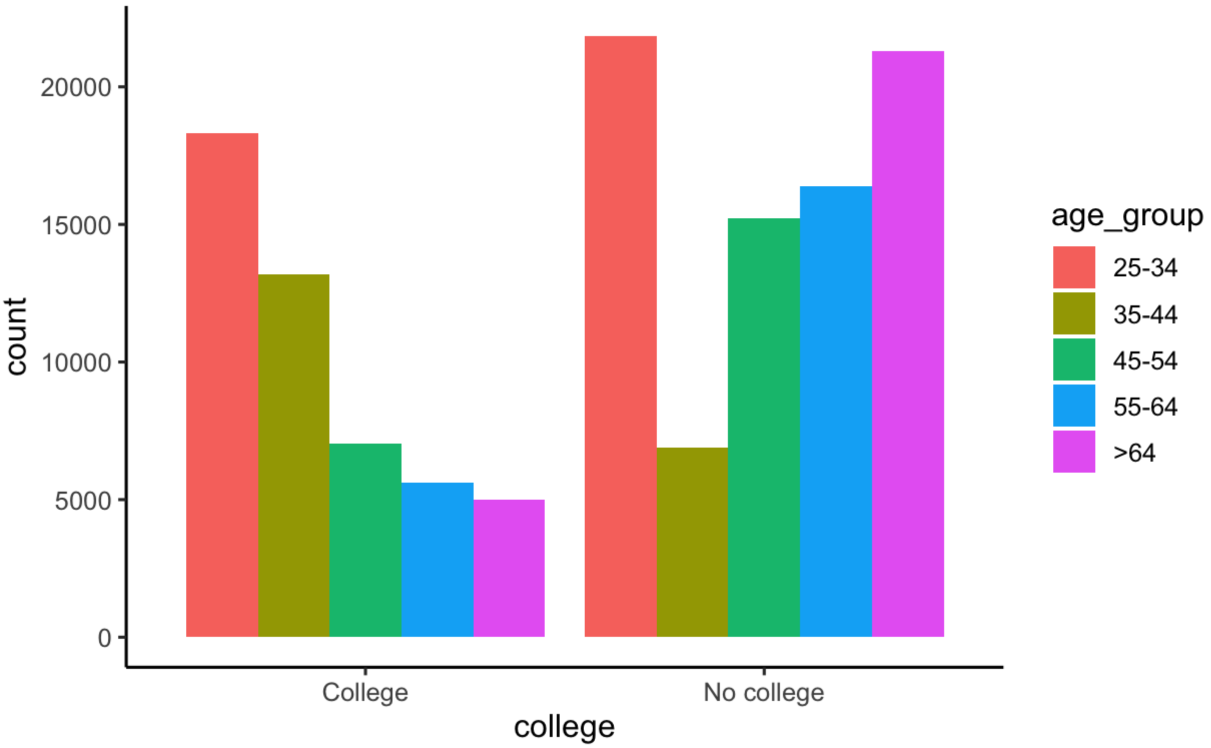



```r
## Examine proportion within each population instead of counts.
ggplot(edu, aes(x = education, y = count, fill = age_group)) + geom_bar(stat = "identity",
    position = "fill")
```

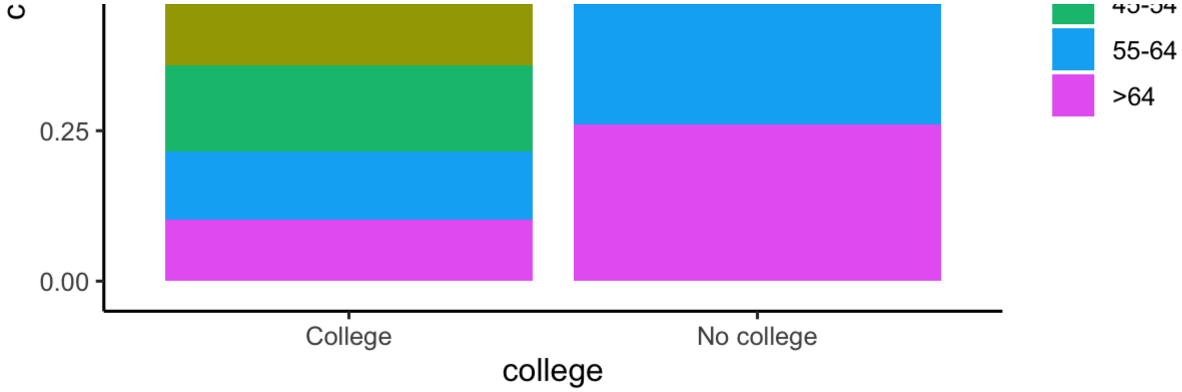





```r
## Limit to look at the category 'finished college'
ggplot(edu_college, aes(x = college, y = count, fill = age_group)) + geom_bar(stat =
        "identity",
    position = position_dodge())
```




```r
## Examine proportion within each population instead of counts.
ggplot(edu_college, aes(x = college, y = count, fill = age_group)) + geom_bar(stat =
        "identity",
    position = "fill")
```



This is census data, there is no sampling and hence no statistical test is appropriate (i.e. no chi-square test).

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

- 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?

```
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")
```

1. It is not appropriate to use relative risk here as the study is retrospective and the participants were enrolled by disease status (heart attack) and not the risk.
2. The odds ratio is

$$\widehat{\mathrm{OR}} = \frac{ad}{cb} = \frac{33 \times 182}{167 \times 18} = 2$$

   We can interpret this as: the odds of being a smoker is 2 times higher for people who have had a heart attack compared to people who have not had a heart attack.

3. $$\mathrm{SE}(\log(\widehat{\mathrm{OR}})) = \sqrt{1/33 + 1/18 + 1/167 + 1/182} = 0.31$$

   so the 95% CI for log odds-ratio is

$$\log(\widehat{\mathrm{OR}}) \pm z^{\star}\ \mathrm{SE}(\log(\widehat{\mathrm{OR}})) = 0.69 \pm 1.96 \times 0.31 \approx (0.086, 1.3)$$

   and the CI for the odds-ratio is there for $(e^{0.086}, e^{1.3}) \approx (1.1, 3.7)$. The “neutral” value for the odds-ratio, 1, does not lie in this CI so there is significant evidence of an association between heart attack and smoking at the 5% level of significance.

```r
x
```

```
           Heart attack: yes Heart attack: no
Smoke: yes                33               18
Smoke: no                167              182
```

```r
y = x[c(2, 1), ]  # rearrange rows as the function is expecting
summary(mosaic::oddsRatio(y))
```

```

Odds Ratio

Proportions
       Prop. 1:  0.4785
       Prop. 2:  0.6471
     Rel. Risk:  1.352

Odds
        Odds 1:  0.9176
        Odds 2:  1.833
    Odds Ratio:  1.998

95 percent confidence interval:
     1.074 < RR < 1.703
     1.084 < OR < 3.683
```

References

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