Problem Set 9 Key

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(epitools)
```

1 Problem 1

Suppose a newly developed statistics class was implemented across the different teaching modalities: asynchronous online, web-based, hybrid, and in-person. A researcher claims that the expected number of students who will fail the class when taught as an asynchronous online course is double that of other modalities, which are assumed to have similar failing rates. 2680 students were randomly assigned to the different class modalities (n=670 each). 90 students failed the asynchronous online course, 45 failed the web-based course, 30 students failed the hybrid course, and 38 failed the in-person course. Use the goodness-of-fit to test the researcher's claim.

a. What are the expected probabilities of failure for each modality? [1pt.]

If p is the probability of failure for other modalities, asynchronous online course is $2p$. The expected probabilities can then be calculated as $2p + p + p + p = 1$, $p = 0.2, 2p = 0.4$.

```r
p <- c(0.4,0.2,0.2,0.2)
```

b. What are the null and alternative hypotheses for this problem? [1pt.]

Null hypothesis: The observed distribution of the number of failed students follows the expected distribution (twice as much for asynchronous online course compared to others).

Alternative hypothesis: The observed distribution of the number of failed students does not follow the expected distribution (twice as much for asynchronous online course compared to others).

c. Implement the goodness-of-fit test. What is the value of the test statistic and the number of degrees of freedom for the corresponding chi-squared distribution? [1 pt.]

```r
ctest <- chisq.test(c(90,45,30,38),p=p)
ctest
```

```
    Chi-squared test for given probabilities

data:  c(90, 45, 30, 38)
X-squared = 4.3645, df = 3, p-value = 0.2247
```

d. Using 0.05 as the significance level, do we reject or fail to reject the null hypothesis? [1pt.]

Fail to reject the null hypothesis.

e. Comment on the sufficiency of evidence against the null hypothesis. [1 pt.]

We have insufficient evidence to claim that the observed distribution of the number of failed students does not follow the expected distribution.

f. Does the chi-squared assumption hold for this problem? [1pt.]

```r
ctest$expected
```

```
[1] 81.2 40.6 40.6 40.6
```

The expected number for each cell is above 5, hence the chi-squared assumption holds.

2 Problem 2

The data set HINTS_smoking.csv contains a subset of the HINTS 6 data set. The columns included are the reported feelings about the participant's household income (variable IncomeFeelings) and current e-cigarette use for those who have history reporting e-cigarette use (variable UseECigNow).

```
smoke <- read.csv("datasets/HINTS_smoking.csv")
glimpse(smoke)
```

```
Rows: 814
Columns: 2
$ IncomeFeelings <chr> "Getting by on present income", "Finding it difficult o~
$ UseECigNow     <chr> "Not at all", "Not at all", "Not at all", "Not at all",~
```

a. Use the test of independence to determine whether there is evidence that income feelings and current e-cigarette use are not independent. What are the null and alternative hypotheses? [1pt.]

Null hypothesis: Income feelings and current e-cigarette use are independent.

Alternative hypothesis: Income feelings and current e-cigarette use are not independent.

b. Implement the test of independence. What is the value of the test statistic and the number of degrees of freedom for the corresponding chi-squared distribution? [1 pt.]

```
ctest <- chisq.test(smoke$IncomeFeelings,smoke$UseECigNow)
ctest
```

```
    Pearson's Chi-squared test

data:  smoke$IncomeFeelings and smoke$UseECigNow
X-squared = 8.4758, df = 6, p-value = 0.2053
```

The test statistic is $\chi^2 = 8.475813$ with 6 degrees of freedom.

c. Using 0.05 as the significance level, do we reject or fail to reject the null hypothesis?

The p-value is 0.205274. We fail to reject the null hypothesis.

d. Comment on the sufficiency of evidence against the null hypothesis. [1pt.]

Insufficient evidence that the income feelings and current e-cigarette use is not independent.

e. Does the chi-squared assumption hold for this problem? Why? [1pt.]

```
ctest$expected
```

```
                                                smoke$UseECigNow
smoke$IncomeFeelings                             Everyday Not at all Some days
  Finding it difficult on present income       18.915233  134.31941 19.765356
  Finding it very difficult on present income   9.074939   64.44226  9.482801
  Getting by on present income                 34.331695  243.79361 35.874693
  Living comfortably on present income         26.678133  189.44472 27.877150
```

The chi-squared assumption holds (expected numbers all greater than 5).

3 Problem 3

Suppose researchers wanted to investigate the prospective careers of graduating seniors from three different high schools. 200 graduating seniors were sampled from each school, and the results of the survey are given below.

```
seniors <- data.frame(School= rep(c("A","B","C"),each=6),
                      Track = rep(c("STEM","Trade","Healthcare","Business", "SS","Others"),t
                      Count=c(57,42,17,21,13,50,45,29,30,30,23,43,50,35,25,34,26,30))
```

```
crosstabs <- xtabs(Count~School+Track,data=seniors)
crosstabs
```

```
      Track
School Business Healthcare Others SS STEM Trade
     A       21         17     50 13   57    42
     B       30         30     43 23   45    29
     C       34         25     30 26   50    35
```

a. Use the test of homogeneity to determine whether the probabilities of pursuing a prospective track is homogeneous across the different schools. What are the null and alternative hypotheses? [1pt.]

Null hypothesis: The marginal probabilities are homogeneous across the different schools.

Alternative hypothesis: The marginal probabilities are not homogeneous across the different schools.

b. Implement the test of independence. What is the value of the test statistic and the number of degrees of freedom for the corresponding chi-squared distribution? [1 pt.]

```
ctest <- chisq.test(crosstabs)
ctest
```

```
    Pearson's Chi-squared test

data:  crosstabs
X-squared = 20.051, df = 10, p-value = 0.02877
```

c. Using 0.05 as the significance level, do we reject or fail to reject the null hypothesis? [1pt.]

Reject the null hypothesis.

d. Comment on the sufficiency of evidence against the null hypothesis. [1pt.]

There is sufficient evidence that the marginal probabilities of being chosen as prospective career tracks are not homogeneous.

e. Does the chi-squared assumption hold for this problem? Why? [1pt.]

```
ctest$expected
```

```
      Track
School Business Healthcare Others       SS     STEM    Trade
     A 28.33333         24     41 20.66667 50.66667 35.33333
     B 28.33333         24     41 20.66667 50.66667 35.33333
     C 28.33333         24     41 20.66667 50.66667 35.33333
```

Chi-squared assumption holds (all expected counts > 5)

4. A randomized controlled trial was used to test the efficacy of a novel intervention to curb excessive recreational cannabis use. 100 participants were randomly assigned equally to two groups (n=50 each group): the treatment group that will receive the newly developed intervention, and a control group that will receive a knowledge-based intervention. The participants received treatment and were evaluated at the midpoint of the study. At this time, 19 participants in the treatment group and 7 participants in the control group reported improvement in controlling the urge to use cannabis recreationally.

```r
rct <- expand.grid(Treatment = c("Control","Intervention"),Outcome=c("Improvement","NoImprove

rct$Count <- c(7,19,50-7,50-19)

crosstabs <- xtabs(Count~Treatment+Outcome,data=rct)
crosstabs
```

```
             Outcome
Treatment      Improvement NoImprovement
  Control                7            43
  Intervention          19            31
```

a. Is this a retrospective study or a prospective study? Why? [2pts.]

Prospective study. The risk was assigned at the start of the study. The two cohorts were followed up to the outcome measurement.

b. Calculate the appropriate measure of association with the 95% confidence interval. [2pts.]

> **Note**
>
> The appropriate measure of association is the relative risk.

```r
epitab(crosstabs,method="riskratio",rev="columns")
```

```
$tab
             Outcome
Treatment      NoImprovement   p0 Improvement   p1 riskratio    lower    upper
  Control                 43 0.86           7 0.14  1.000000       NA       NA
  Intervention            31 0.62          19 0.38  2.714286 1.253168 5.878978
             Outcome
Treatment         p.value
  Control              NA
  Intervention 0.01129646

$measure
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "fisher.exact"
```

The risk ratio is 2.71, with 95% CI: (1.24, 5.88).

5. Researchers studied the effect of cranberry juice in the treatment and prevention of Helicobacter pylori infection in mice. Researchers compared treatment with cranberry juice to "triple therapy (amoxicillin, bismuth subcitrate, and metronidazole) in mice infected with Helicobacter pylori. After 4 weeks, they examined the mice to determine the frequency of eradication of the bacterium in the two treatment groups.

```
rct <- expand.grid(Treatment = c("Triple","Cranberry"),Outcome=c("Yes","No"))

rct$Count <- c(8,2,2,8)

crosstabs <- xtabs(Count~Treatment+Outcome,data=rct)
crosstabs
```

```
           Outcome
Treatment   Yes No
  Triple      8  2
  Cranberry   2  8
```

   a. Calculate the expected values based on the assumption of homogeneity. [2pts.]

```
ctest <- chisq.test(crosstabs)
ctest$expected
```

```
           Outcome
Treatment   Yes No
  Triple      5  5
  Cranberry   5  5
```

   b. Does the chi-squared assumption hold? [1pt.]

The expected values of each cell is equal to 5, which is a boundary case. The expected values are not greater than 5, hence the chi-squared distribution assumption might not hold.

   c. Perform the appropriate test to test for an association between treatment and the outcome. [2pts.] Hint: Exact tests are used for small expected counts.

```
fishtest <- fisher.test(crosstabs)
fishtest
```

```
    Fisher's Exact Test for Count Data

data:  crosstabs
p-value = 0.02301
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   1.309537 239.395560
sample estimates:
odds ratio
  13.25038
```

The p-value is 0.0230141, which is lower than 0.05. We reject the null hypothesis. There is sufficient evidence of an association between treatment and eradication of *H. pylori*.