STA 3100 Homework 7
Due on April 25, 2023 at 11:59 p.m.

Instructions:

- Submit the following two files to the relevant dropbox on course website.

  1. **One single pdf file of your report.** This file should be named HWx_*Last*.pdf, where x is the homework number and *Last* is your last name (e.g., HW1_Lee.pdf). Show your steps, code, and output in your report so that the grader can see your work without actually running your code. You do not need to explain your code line by line, but you should provide enough details so that the grader can understand the flow of your code.
  2. **One single R file of your code.** This file should be named HWx_*Last*.R, where x and *Last* are the same as before. This file is to allow the grader to verify your work if necessary. The grader should be able to run your code without modification except (possibly) changing the paths that you use for filenames, etc.

  In summary, your report should stand alone!

- **You are not allowed to use any non-Base R package unless otherwise specified.**

- **Do not use R comments to answer a theoretical question. Handwrite or type your answer clearly with appropriate statistical notations.** For example, do not write something like

  - `# H1: mu != mu0;`
  - `# p-value is pnorm(zstar).`

  Instead, handwrite or type

  - H_1: $\mu \neq \mu_0$;
  - p-value is $P(Z < z^*)$,

  with all symbols defined clearly.

- **Problems should appear in the order that they were assigned.**

- **Failing to follow the instructions above may result in a deduction of 50% or more credit.**

Assignment:

1. Suppose $Y_{1i} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_{2j} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_2, \sigma_2^2), i = 1, \ldots, n_1, j = 1, \ldots, n_2$, where Y_{1i}'s and Y_{2j}'s are all independent. We would like to conduct an F-test to see if the variances are equal (i.e., $\sigma_1^2 = \sigma_2^2$). We would also like to construct a $100(1 - \alpha)\%$ confidence interval for σ_1^2/σ_2^2. Recall the test and confidence interval from your introductory statistics course (e.g., STA 3032), which were also reviewed in this course. Define your notation clearly and answer the following questions.

   (a) State the null and alternative hypotheses. Note that there are three possible alternatives.

   (b) Write the test statistic and its null distribution.

   (c) For each alternative, write the p-value and explain how to draw a conclusion based on it at the significance level α.

   (d) Construct a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2. You do not need to consider a one-sided confidence bound.

   (e) Write a function to conduct the tests and construct the interval above. No credit will be given if any built-in function related to the variance test is used. You may use the four Base R functions for a given distribution (e.g., `d/p/q/rf()`). Your function should perform as follows.

      i. The function takes the arguments:
      `y1`, `y2`, `alt = "two-sided"`, `lev = 0.95`,
      where the equality indicates the default value.

         - The arguments `y1` and `y2` are the two samples.
         - The argument `alt` is the alternative hypothesis whose two other possible values are `"less"` and `"greater"`.
         - The argument `lev` is the confidence level $1 - \alpha$.

      ii. The function returns an R list containing the test statistic, p-value, and confidence interval.

      iii. Inside the function, two Shapiro-Wilk tests of normality are conducted separately for the two samples (note the normality assumption at the beginning of the problem). If one or both p-values are less than 0.05, a warning message is printed out explaining the situation. Regardless, the function performs part (ii).

   (f) Use your function above to solve the following problem.
   Problem: The following data represent the running times of films produced by two motion-picture companies.

Company	Time (minutes)									
I	139	131	147	108	122	129	140	144	86	104
II	169	239	120	124	99	113	96	125		

Test the hypothesis that the variances for the running times of films produced by Company I and Company II are equal against the alternative that they are not equal. Draw a conclusion based on a p-value at the 0.01 significance level. Construct and interpret a 99% confidence interval for the ratio of the variances.

2. Suppose $Y_{1i} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_{2j} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_2, \sigma_2^2), i = 1, \ldots, n_1, j = 1, \ldots, n_2$, where Y_{1i}'s and Y_{2j}'s are all independent. In Homework 6, we wrote a function to test $H_0 : \mu_1 = \mu_2$ and construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ assuming the variances σ_1^2 and σ_2^2 were unknown but equal. In this problem, we modify the function to incorporate the case where the variances are unknown and unequal. Define your notation clearly and answer the following questions. For parts (a) and (b), assume σ_1^2 and σ_2^2 are unknown and unequal.

   (a) Write the test statistic and its null distribution. Specify the null distribution degrees of freedom using the Satterthwaite approximation discussed in class.

   (b) Construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$. You do not need to consider a one-sided confidence bound.

   (c) Modify the two-sample t-test function you wrote in Homework 6 so that the function works whether or not the unknown variances are equal. No credit will be given if any built-in function related to the t-test is used. You may use the four Base R functions for a given distribution (e.g., `d/p/q/rt()`). Your function should perform as follows.

      i. The function takes the arguments:
      `y1`, `y2`, `alt = "two-sided"`, `lev = 0.95`,
      where the equality indicates the default value.

         - The arguments `y1` and `y2` are the two samples.
         - The argument `alt` is the alternative hypothesis whose two other possible values are `"less"` and `"greater"`.
         - The argument `lev` is the confidence level $1 - \alpha$.

      ii. The function first conducts a variance test using the function you wrote in Problem 1. If the variance test p-value is greater than 0.05, the function assumes $\sigma_1^2 = \sigma_2^2$; otherwise, it assumes $\sigma_1^2 \neq \sigma_2^2$. Based on that assumption, the function tests $H_0 : \mu_1 = \mu_2$ and constructs a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

      iii. The function returns an R list containing the test name, test statistic, p-value, and confidence interval, where the test name takes one of the following two values

         A. `"two-sample t-test with unknown but equal variances"`;

         B. `"two-sample t-test with unknown and unequal variances"`.

      iv. The function does not return the variance test result.

   (d) Use your new t-test function above to repeat part (g) of Problem 6 in Homework 6. Are the results same as before? Why or why not?

3. Recall the chi-squared test of independence for two categorical variables from your introductory statistics course (e.g., STA 3032), which was also reviewed in this course. Let X and Y be the two categorical variables, and let $\mathbf{O} = [o_{ij}], i = 1, \ldots, r, j = 1, \ldots, c$, be the $r \times c$ contingency table that cross-tabulates the $n_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij}$ observations by X and Y. That is, X and Y have r and c levels, respectively, and o_{ij} is the observed count for the (i, j)th cell. Finally, let $n_{i.} = \sum_{j=1}^{c} o_{ij}$ and $n_{.j} = \sum_{i=1}^{r} o_{ij}$ be the ith row total and the jth column total, respectively. Define your notation clearly and answer the following questions.

   (a) State the null and alternative hypotheses.

   (b) Explain how to calculate the expected count e_{ij} for the (i, j)th cell under the null.

   (c) Write the test statistic and its null distribution.

   (d) Explain why a large value of the test statistic provides evidence against the null.

   (e) Write the p-value and explain how to draw a conclusion based on it at the significance level α.

   (f) Write a function to conduct the chi-squared test of independence. No credit will be given if any built-in function related to the chi-squared test is used. Also, you are not allowed to use the `outer()` function or the `%o%` operator. You may use the four Base R functions for a given distribution (e.g., `d/p/q/rchisq()`). Your function should perform as follows.

      i. The function takes the arguments:
      `dat`, `res.type = "pearson"`,
      where the equality indicates the default value.

      - The argument `dat` is an R matrix of the $r \times c$ contingency table.
      - The argument `res.type` specifies the type of the residuals whose other possible value is `"std"`. Here, the type `"pearson"` indicates the Pearson residual

      $$r_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

      and the type `"std"` indicates the standardized residual

      $$r_{ij}^{*} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}(1 - n_{i.}/n_{..})(1 - n_{.j}/n_{..})}}.$$

      ii. The function does not use continuity correction.

      iii. The function returns an R list containing the test statistic, p-value, expected counts, residual type, and the residuals of that type, where the expected counts and the residuals are stored in two $r \times c$ matrices.

      iv. Recall that the chi-squared approximation requires each expected count to be at least 5. If any expected count is less than 5, the function prints out a warning message informing potential inaccuracy of the approximation. Regardless, the function performs part (iii).

(g) Use your function above to solve the following problem.
Problem: A survey asked a random sample of 1397 U.S. adults about their education level and marital status. The results are summarized below.

	Marital Status		
Education	Single/Widowed	Married	Divorced
Elementary	6	18	13
Secondary	115	256	136
College	256	442	155

Test if one's marital status is independent of their education level. Draw a conclusion based on a p-value at the 0.05 significance level. Calculate the standardized residuals.

4. Suppose we want to generate a random sample from the distribution with pdf

$$f(x) = \begin{cases} 6x(1-x), & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

To that end, we generate a random number y from a distribution with pdf $g(y)$ and another random number u from Unif$(0, 1)$. Then we take y as an observation from $f(x)$ if

$$u < \frac{f(y)}{c \cdot g(y)},$$

where c is a constant; otherwise, we discard y and generate new y and u, and repeat.

(a) Three choices of $g(y)$ and c are given below. In each case, generate a random sample of size $n = 1000$ from $f(x)$ using the method above with a while loop. Why is a while loop appropriate here? Report the efficiency defined as the sample size divided by the total number of y generated. Which $g(y)$ seems to be the most efficient? Which one is the least efficient?

i. $g(y) = \text{Unif}(0, 1)$ and $c = 1.5$.

ii. $g(y) = \text{N}(0.5, 0.05)$ and $c = 2$.

iii. $g(y) = t_3$ (i.e., t with 3 degrees of freedom) and $c = 5$.

(b) You should now have three random samples of size n. Produce three histograms of the samples with y-axis ranging from 0 to 4. Add the pdf $f(x)$ in red and the (scaled) pdf $c \cdot g(y)$ in blue to each histogram (you may not see the overall shape of $c \cdot g(y)$. why?). Split the plot window into two rows and three columns, and arrange the histograms in the first row.

(c) The file true_quantiles.dat in the Datasets folder under Files on course website contains n theoretical quantiles of $f(x)$ corresponding to probabilities 0.0005, 0.0015, ..., 0.9985, 0.9995. Read the quantiles into R. For each of the three random samples, produce a Q-Q plot of sample quantiles vs. theoretical quantiles. Arrange the Q-Q plots in the second row of the plot in part (b) (i.e., below the histograms). Which sample(s) seem to agree with $f(x)$? Explain.