Due: April 11th, 2024,10pm
Problems Due: 2,3,5,8

1. Shyamala is interested in studying average years of schooling in India. They hypothesized that the mean years of school for people 18 years old or above is higher than 8.5 years. They drew a sample of 300000 from 2011 census and find out that the mean number of years of schooling for the sample population is 8.515, with a SD of 4.5. We wish to test

$$H_0 : \mu \leq 8.5 \text{ versus } H_A : \mu > 8.5$$

   (a) Without performing any analysis, based on your intuition would you reject the null hypothessis ?

   (b) Compute the T-statistic.

   (c) Find the p-value.

   (d) Describe the meaning of p-value and what inference can you draw from it if your level of significance is 0.005

2. *Two sample test of means for paired data* Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two samples, that are paired. Then the test Statistic is given by:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{\frac{S}{\sqrt{n}}} \quad \text{where} \quad S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(z_i - \bar{z})^2\right) \quad \text{with } z_i = x_i - y_i \text{ with } T \sim t_{n-1}$$

   (a) Write a function called `pairedttest` that performs the above test for paired data x and y.

   (b) Consider the `shoes` dataset from the package `MASS`. It contains shoe wear data. There is a list of two vectors, giving the wear of shoes of materials A and B for one foot each of ten boys.

      i. Using the function `pairedttest` perform the paired t-test on the above data to see if the two types of shoes have different mean wear amounts.

      ii. Verify the same using the inbuilt function `t.test(A,B, paired=TRUE)`.

      iii. If you were to use the command `t.test` without the `paired=TRUE` command then can you still perform the above test ?

3. Take the dataset `heartrate` from the package `UsingR`.

   (a) Make a scatter plot of the variables `maxrate` and `age`. Using the inbuilt `lm` function plot the regression line of `maxrate` $= \beta_0 + \beta_1$`age` on the same plot.

   (b) Using the inbuilt `resid` function compute the residuals at each point. Using the inbuilt `fitted` function compute the fitted values at each point. Verify that `resid` = actual values of data - `fitted`.

   (c) Using the generic `predict` function or otherwise, from the above line find the `maxrate` for `age` $= (12, 33, 45, 76)$.

   (d) Write a simple `function` to test the null hypothesis that $\beta_1 = 0$ versus the alternative that $\beta_1 \neq 0$. Decide if the data has evidence to reject the null hypothesis.

   (e) From the estimate of $\hat{\beta}_1$ obtained above find a 95% confidence interval for β_1.

   (f) Do the following plots:

      i. Scatter plot of residuals versus fitted values.

      ii. Scatter plot of $\sqrt{\text{standardised residuals}}$ versus fitted values.

      iii. A NormaaQ-Q plot for the residuals.

4. The student body at an undergraduate university is 20% Masters, 24% third years, 26% Second year, and 30% first year students. Suppose a researcher takes a sample of 50 such students. Within the sample there are 13 Masters, 16 Third years, 10 Second years, and 11 First years. The researcher claims that his sampling procedure should have produced independent selections from the student body, with each student equally likely to be selected. Is this a plausible claim given the observed results?

5. In an experiment in breeding plants, a geneticist has obtained 120 brown wrinkled seeds, 48 brown round seeds, 36 white wrinkled seeds and 13 white round seeds. Theory predicts that these types of mice should be obtained in the ratios 9:3:3:1. Is the theory a valid hypothesis ?

6. Save data collected in class from the random number tables as `rnt.csv`. Write a function in R called `chisquarefit` to perform the chi-square goodness of fit test to decide if the numbers drawn from the table were truly random.

7. From the Talent experiment data:

   (a) Save the data, on the trial at which the first success was obtained, collected in class from the talent experiment as `firstsuccess.csv`. Write a function `mlefs` to compute the likelihood function for the probability of success. Then using the inbuilt `optim` function find the m.l.e for the probability of success.

   (b) Save the data, on the number of successes in 10 trials, collected in class from the talent experiment as `numsuccess.csv` data. Write a function `mlens` to compute the likelihood function for the probability of success. Then using the inbuilt `optim` function find the m.l.e for the probability of success.

   (c) Can you perform a test whether your groups Probability of success is the same as the Probability of success for the class ?

8. Read in `Annualtemp` data from the shared dropbox folder.

   (a) Plot Temperature vs CO2. In this plot, using `points` mark the mean of both variables as a special point with red color. Using the `abline`, `v` and `h` plot two lines, horizontal and vertical respectively at each of the means.

   (b) Compute the correlation between Temperature and CO2.

   (c) Plot the linear regression line.

   (d) Test whether the slope is 0 or not.

9. In the `ToothGrowth` data set observe that there are lot of repetitions in the `ToothGrowth$dose` variable.

   (a) Find the correlation coefficient between `ToothGrowth$len` and `ToothGrowth$dose`.

   (b) Plot `ToothGrowth$len` versus `ToothGrowth$dose`.

   (c) Using `split` command find the group means for each duplication of `ToothGrowth$dose`.

   (d) Find the correlation coefficient between group means of `ToothGrowth$len` and the appropriate `ToothGrowth$dose`.

Book-Keeping Exercises[1]

1. Let X be a random variable with finite range $\{c_1, c_2\}$ for which $P(X = c_j) = p_j > 0$ for $j = 1, 2$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with distribution X and let $Y_j = |\{j : X_j = c_j\}|$. Then $\chi^2 = \sum_{j=1}^{2} \frac{(Y_j - np_j)^2}{np_j}$ has the same distribution as $(\frac{Z - E[Z]}{SD[Z]})^2$ where $Z \sim \text{Binomial}(n, p_1)$

2. Let X be a random variable with finite range $\{c_1, c_2\}$ for which $P(X = c_j) = p_j > 0$ for $j = 1, 2$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with distribution X and let $Y_j = |\{j : X_j = c_j\}|$. Define

$$X^2 = \sum_{j=1}^{2} \frac{(Y_j - np_j)^2}{np_j}.$$

   Let $F_n : [0, \infty) \to [0, 1]$ be the distribution function of X^2.

   Show that there is a differentiable $F : \mathbb{R} \to [0, 1]$ such that $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$ and find F'.

3. χ^2 *goodness of fit test for continuous random variables*

   (a) Generate 1000 observations from uniform $[0, 1]$ in `R`.

   (b) Compute Y_j, the number of observations in $[\frac{j-1}{10}, \frac{j}{10}]$ with $j \in \{0, 1, \ldots, 10\}$.

   (c) Perform a χ^2 goodness of fit test on the null hypothesis being $p_1 = p_2 = \ldots p_{10} = \frac{1}{10}$ verus the alternative hypothesis being $(p_1, p_2, \ldots, p_{10}) \neq (\frac{1}{10}, \ldots, \frac{1}{10})$ for 5% level of significance.

   Do the above when 10 is replaced by $25, 50$, and 100.

4. Let $\{T_i : i \geq 1\}$ be Exponential (λ) random variables. Define $X := \max\{j : \sum_{k=1}^{j} T_k \leq 1\}$

   (a) Show that for $l \in \mathbb{N}$,

$$\frac{\lambda^l}{l - 1!} \int_x^\infty z^{l-1} e^{-\lambda z} dz = \sum_{m=0}^{l-1} e^{-\lambda x} \frac{(\lambda x)^m}{m!}$$

   (b) Show that $n \geq 0$, $P(X > n) = P(\sum_{k=1}^{n+1} T_k \leq 1)$

   (c) Suppose we have a random number generator that produces a sample U from a Uniform$(0, 1)$. Describe a simulation procedure from the above two parts that will generate a sample from Poisson (λ) using U.

5. Write a function called `NBsim` to simulate a sample from NegativeBinomial (n, p) from a U generated by using the `runif` command.

6. Write a function called `Csim` to simulate a sample from the standard Cauchy distribution from a U generated by using the `runif` command.

7. Execute the following `R` commands.

```
> x = seq(-10,10,by=1/10)
> y = 3*x^2 - 0.0001*x +7
> plot(y~x,pch=16, cex=0.2, col="red")
> abline(lm(y~x), col="blue")
```

   (a) Explain the reasons behind the poor simple regression line in contrast to the clear relationship between x and y

   (b) Using the `optim` command can you find estimates for $\beta_0, \beta_1, \beta_2$ that minimize $\sum_{i=1}^{100}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)$

[1] Questions are from material within the course syllabus and you may be tested on these.