MATH255 Autumn 2023 Computer Lab - Week 12

Note: Question 1(b) is your Lab Preparation exercise for this week. Your answers must be submitted as a pdf document before the start of your lab.

Key Results from Lectures

- $E(X) = \int_{-\infty}^{\infty} xf(x)dx, E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$ — Expected Value, Variance
- $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(z)dz$ — Cumulative Distribution, Quartiles
- $f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{else.} \end{cases}, \mu = \frac{b-a}{2}, \sigma^2 = \frac{(b-a)^2}{12},$ `dunif, punif` — Uniform Distribution
- $f(x) = \lambda e^{-\lambda x}, F(x) = 1 - e^{-\lambda x}, \mu = \frac{1}{\lambda},$ `dexp, pexp` — Exponential Distribution
- $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}},$ `dnorm, pnorm` — z-critical Value
- $t = \frac{\bar{x}-\mu}{s\sqrt{n}},$ `dt, pt` — t-critical Value
- $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ — Confidence Interval, σ known
- $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ — Confidence Interval, σ unknown

1. Download the Uniform.csv file from Moodle and upload it to R. It contains 25 independent samples $X_1, \ldots, X_{25}$, each of size $n = 25$, from a continuous uniform distribution between 0 and 1, denoted by $U[0, 1]$.

   (a) Add three more columns to the spreadsheet, called MEAN, VAR and SD, that contain the mean, variance and standard deviation of each sample variable. The syntax of a for loop in R is the following. `for (i in 1:25) { ... }`

   (b) What are the population mean, variance and standard deviation of a $U[0, 1]$ variable? Then what are the expectation and standard error of the sample mean for a sample of size 25?

   (c) Calculate the mean and standard deviation of your MEAN variable. How do these compare to your answers in (b)?

   (d) Plot a histogram of MEAN. Does the Central Limit Theorem idea of approximate normal distribution seem to apply here? How would you expect the distribution to change if the sample size were increased?

   (e) Calculate the mean of your variables VAR and SD. How do these compare to your answers in (b)? What will happen to these calculations as the sample size increases?

2. Define a vector x (`seq` command) that has values in numerical order from 0 to 5, incrementing by 0.01 (so 501 points). Generate a variable y that is an exponential distribution of x. Plot (x, y) and verify that the graph is an exponential distribution.

   (a) Take 100 random samples of size 5 from y (`sample` command) and record the mean of each sample. Plot a histogram of the means.

(b) Repeat for 100 samples of size 20, then 100 samples of size 100. How do the histograms change? Is this what you expect to happen?

3. The `mtcars` data set in R contains statistics about 32 different models of race car.

   (a) Take a random sample of size 8 from the Horsepower column. Calculate the sample mean and sample standard deviation. Construct a 95% confidence interval for the mean horsepower of all the cars, using only the information you calculated and an appropriate distribution table value. Repeat for another random sample of size 12 and note any differences/similarities between the two.

   (b) Calculate the population standard deviation and use it to find 95% confidence intervals for your two samples in (a). Why are these intervals different than the ones you found in (a)? Which intervals do you think are better, if any?

   (c) Calculate the population mean and see if it is within your intervals. If it is outside one or more of your intervals, why did that happen?