

Question 1

Suppose that X_1, X_2, \dots, X_n are independent random variables that follow a $N(\mu, \sigma^2)$ distribution, and define $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X} - X_i)^2$, as usual. Show that the random variable T , where

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

can be written in the form

$$T = \frac{U}{\sqrt{V/p}},$$

where

- $U \sim N(0, 1)$,
- p is some function of n ,
- $V \sim \chi_p^2$, the chi-squared distribution with p degrees of freedom,
- U and V are independent random variables.

Question 2

Suppose the following 11 values are the transaction amounts (in £) of online purchases for a particular credit card customer in a given month.

$$45, 81, 52, 23, 147, 92, 76, 124, 287, 103, 65$$

Tukey's criterion states that, given the lower quartile $q_{0.25}$, the upper quartile $q_{0.75}$ and the interquartile range IQR, if a value x is either $x < q_{0.25} - k\text{IQR}$ or $x > q_{0.75} + k\text{IQR}$, for $k = 1.5$, then x is considered to be an outlier.

- (a) Compute the lower and upper quartiles, and the interquartile range for this dataset.
- (b) According to Tukey's criterion, are any of these transaction amounts outliers?
- (c) If any of the transactions is an outlier, would you take any action? What could be the consequences of
 - (i) inaction (doing nothing) or (ii) taking action (preventing the transaction from going through)?
- (d) If you were designing your own fraud detector for this customer (not using Tukey's criterion) for the next month, how high would a value need to be for you to decide that a value is anomalous and potentially fraudulent? In other words, at what value would you set the threshold?

Question 3 (R question)

It is suggested that the following question is done in an R Markdown document.

- (a) Use `dnorm` to plot the probability density function of the standard normal random distribution on the interval $[-4, 4]$.

Hint: Use the `seq` function to generate 1000 evenly spaced points on the interval $[-4, 4]$.

- (b) Use `dgamma` to plot the probability density function of a $\Gamma(2, 0.5)$ random variable on the interval $[0, 20]$. Note that we are using the shape/rate parametrisation here, i.e. $\alpha = 2$ is the shape and $\beta = 0.5$ is the rate.

- (c) Now do the following:

- (i) For $X_1, X_2, \dots, X_n \sim \Gamma(\alpha, \beta)$, use R to sample observations x_1, x_2, \dots, x_n , where $n = 1000$ and $\alpha = 2$ and $\beta = 0.5$.
- (ii) From these x_1, x_2, \dots, x_n values, compute the the standardised z_1, z_2, \dots, z_n , where

$$z_i = \frac{x_i - \mathbb{E}[X_i]}{\sqrt{\text{Var}[X_i]}}.$$

Hint: For $X \sim \Gamma(\alpha, \beta)$, $\mathbb{E}[X] = \frac{\alpha}{\beta}$ and $\text{Var}[X] = \frac{\alpha}{\beta^2}$.

- (iii) Compute the weighted sum

$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i.$$

(Note the square root in the fraction $1/\sqrt{n}$; this is **not** the sample mean.)

- (iv) Repeat steps (a) to (c) t times (using a loop), and save the resulting sums S_1, S_2, \dots, S_t to a vector \mathbf{S} . It is suggested that t is set to $t = 10,000$.
- (v) Plot a histogram of the values S_1, S_2, \dots, S_t . In the `hist` function, set the parameters `freq=FALSE` and `breaks=30`.
- (vi) Does this histogram look familiar? Use the `lines` function in R to plot the probability density function of an appropriate distribution over the histogram.