Department of Mathematics and Statistics

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

CSE Exercises – Week 2

① In this exercise, we find out why we divide by $(n-1)$ instead of n in the definition of the sample variance given in equation (5.4).

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with population mean, $E(X_1)$, and population variance, $V(X_1)$. Recall that the definition of the sample variance is

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$$

(a) Show that $E(S_n^2) = V(X_1)$. This says that the expectation of the sample variance is equal to the population variance. In other words, when the sample variance is used to estimate the population variance, it will be equal, "on average", to the population variance. An estimator that is equal, "on average", to the quantity that it is estimating is described as "unbiased" because the bias, defined as $E(S_n^2) - V(X_1)$, is zero. Thus, dividing by $(n-1)$ makes the sample variance unbiased. Part (b) reinforces this.

(b) Now consider the following alternative definition for sample variance:

$$T_n^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$$

Show that T_n^2 is a biased (i.e. not unbiased) estimator for the population variance, and find the bias.

(c) Now we perform a simple Matlab experiment that illustrates Parts (a) and (b) computationally.

Do the following:

(i) Generate $n=10$ Uniform$(0,1)$ sample values.
(ii) Compute S_n^2 and T_n^2 and store them. Recall that you can compute S_n^2 using the "var" built-in function in Matlab. In fact, you can use the same function to compute T_n^2 by specifying a second input of 1, for example, var(X, 1).
(iii) Repeat steps (i) and (ii) 10,000 times.
(iv) Compute the sample mean of the 10,000 S_n^2 values; this is an estimate of $E(S_n^2)$.
(v) Compute the sample mean of the 10,000 T_n^2 values; this is an estimate of $E(T_n^2)$.
(vi) Compare the estimates from steps (iv) and (v) with the Uniform$(0,1)$ population variance. Comment on your results.

(2) Revisiting Exercise 4(b) from week 1:

(a) Write down the Uniform $(-1, 1)$ quantile function and plot it.

(b) Find the median, first quartile and third quartile.

(3) Revisiting Exercise 6 from week 1:

(a) Write down the quantile function and plot it.

(b) Find the median, first quartile and third quartile.

(4) Revisiting Exercise 7 from week 1:

(a) Write down the quantile function and plot it.

(b) Find the median, first quartile and third quartile.

(5) Revisiting Exercise 8 from week 1:

(a) Write down the quantile function and plot it.

(b) Find the median, first quartile and third quartile.

(6) Let $a, b \in \mathbb{R}$ such that $a < b$, and let U be a Uniform $(0,1)$ random variable. If

$$V = a + (b-a)U, \qquad (*)$$

then it can be shown mathematically (later on) that V is Uniform(a,b). In this exercise, we will demonstrate this computationally.

(a) Do the following:

(i) Generate $n = 10{,}000$ Uniform$(0,1)$ random values, $U_1, \ldots, U_n$, and store them.
(ii) Use the values from step (i) together with $(*)$ to get $V_1, \ldots, V_n$ having a Uniform$(-1, 1)$ distribution.
(iii) Plot a density histogram for $V_1, \ldots, V_n$.
(iv) Use the values from step (i) together with $(*)$ to get $W_1, \ldots, W_n$ having a Uniform $(0, \frac{1}{2})$ distribution
(v) Plot a density histogram for $W_1, \ldots, W_n$.

(b) Now suppose that X and Y are independent Uniform $(0, \frac{1}{2})$ random variables. Let $Z = X + Y$.

(i) What is the support (set of values of a random variable with non-zero PDF) of Z?

(ii) What do you think is the shape of the PDF of Z? Provide a sketch right now; it does not matter if it turns out to be incorrect.

(iii) Generate $X_1, \ldots, X_{10000} \sim \text{Uniform}(0, \frac{1}{2})$ and $Y_1, \ldots, Y_{10000} \sim \text{Uniform}(0, \frac{1}{2})$ and compute $Z_i = X_i + Y_i$, $i = 1, \ldots, 10000$. Plot a density histogram for $Z_1, \ldots, Z_{10000}$. Comment on the shape of the histogram as an estimate of the PDF of Z. Have you seen this PDF shape before? If so, where?

The message of this exercise is that when two (or more) independent random variables having the same distribution are added, the resulting random variable generally does <u>not</u> have the same type of distribution. There are, however, special distributions for which the sum does have the same type of distribution. The best known example is the normal distribution (which you must have encountered before but will be defined formally later on). If X and Y are independent normal random variables, then $Z = X + Y$ is also a normal random variable (but with different parameter values from the normal distribution for X and Y).

Solutions

① Given $X_1, \dots, X_n$ iid with mean $E(X_1)$ and variance $V(X_1)$.

(a) $$S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$E(S_n^2) = E\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2\right]$$
$$= \frac{1}{n-1} \sum_{i=1}^{n} E\left[(X_i - \bar{X}_n)^2\right].$$

Now look at the expectation within the sum:

$$E\left[(X_i - \bar{X}_n)^2\right]$$
$$= E\left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right)$$
$$= E(X_i^2) - 2E(X_i\bar{X}_n) + E(\bar{X}_n)^2$$
$$= E(X_1^2) - 2E(X_i\bar{X}_n) + E(\bar{X}_n)^2 \qquad (*)$$

since $X_1, \dots, X_n$ are iid

The second expectation in $(*)$ is

$$E(X_i\bar{X}_n) = E\left[X_i\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)\right]$$
$$= E\left(\frac{1}{n}\sum_{j=1}^{n} X_iX_j\right)$$
$$= \frac{1}{n} E\left(\sum_{j=1}^{n} X_iX_j\right)$$
$$= \frac{1}{n} E\left(X_i^2 + \sum_{\substack{j=1 \\ j\neq i}}^{n} X_iX_j\right)$$

$$= \frac{1}{n}\Big[E(X_i^2) + \sum_{\substack{j=1 \\ j\neq i}}^{n} E(X_i X_j) \Big]$$

$$= \frac{1}{n}\Big[E(X_i^2) + \sum_{\substack{j=1 \\ j\neq i}}^{n} E(X_i) E(X_j) \Big]$$

since X_i and X_j are independent when $i \neq j$

$$= \frac{1}{n}\big[E(X_1^2) + (n-1)\, E(X_1)^2 \big].$$

The third expectation in $(*)$ is

$$E(\overline{X}_n^2) = E\Big[\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)\Big(\frac{1}{n}\sum_{j=1}^{n} X_j\Big) \Big]$$

$$= \frac{1}{n^2} E\Big[\Big(\sum_{i=1}^{n} X_i\Big)\Big(\sum_{j=1}^{n} X_j\Big) \Big]$$

$$= \frac{1}{n^2} E\Big(\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} X_i X_j \Big)$$

$$= \frac{1}{n^2}\Big[\sum_{i=1}^{n} E(X_i^2) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} E(X_i X_j) \Big]$$

$$= \frac{1}{n^2}\big[n\, E(X_1^2) + n(n-1)\, E(X_1)^2 \big].$$

Therefore,

$$E(S_n^2) = \frac{1}{n-1}\sum_{i=1}^{n} E\big[(X_i - \overline{X}_n)^2\big]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \Big[E(X_1^2) - \frac{2}{n}E(X_1^2) - 2\Big(\frac{n-1}{n}\Big)E(X_1)^2 + \frac{1}{n}E(X_1^2) + \Big(\frac{n-1}{n}\Big)E(X_1)^2 \Big]$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left[\left(\frac{n-1}{n}\right) E(X_1^2) - \left(\frac{n-1}{n}\right) E(X_1)^2 \right]$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{n-1}{n}\right) \left[E(X_1^2) - E(X_1)^2 \right]$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{n-1}{n}\right) V(X_1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} V(X_1)$$

$$= \frac{1}{n} \cdot n V(X_1) = V(X_1) .$$

(b) $\quad T_n := \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$

Notice that $T_n = \left(\frac{n-1}{n}\right) S_n$, and so

$$E(T_n) = E\left[\left(\frac{n-1}{n}\right) S_n\right]$$

$$= \left(\frac{n-1}{n}\right) E(S_n)$$

$$= \left(\frac{n-1}{n}\right) V(X_1) .$$

Hence, $E(T_n) \neq V(X_1)$ and so T_n is a biased estimator for $V(X_1)$. The bias is

$$E(T_n) - V(X_1) = \left(\frac{n-1}{n}\right) V(X_1) - V(X_1)$$

$$= - \frac{V(X_1)}{n} .$$

Hence, there is a negative bias, which means that T_n will tend to under-estimate $V(X_1)$ "on average".

(c) The sample mean of the 10,000 S_n values is an estimate of $E(S_n^2)$. I got

$$E(S_n^2) \approx 0.0832 .$$

Likewise, the sample mean of the 10,000 T_n values is an estimate of $E(T_n^2)$. I got

$$E(T_n^2) \approx 0.0748 .$$

We know that the population variance of the Uniform (0, 1) is

$$\frac{1}{12} \approx 0.0833 ,$$

which is close to the estimated $E(S_n^2)$ but different from the estimated $E(T_n^2)$, as expected.

Notice that the estimated $E(T_n^2)$ is smaller than the population variance, thus demonstrating the under-estimation predicted by the theory. Since the sample size $n = 10$, the actual bias of T_n^2 is

$$-\left(\frac{1}{10}\right)\left(\frac{1}{12}\right) \approx -0.0083 .$$

The estimated bias is

$$0.0748 - 0.0833 \approx -0.0085 ,$$

which is close.

(2) Since we have a continuous distribution here, we can find the quantile function by finding the inverse of the CDF. From week 1 exercise 4(b), the Uniform(a, b) CDF is, for $x \in [a, b]$,

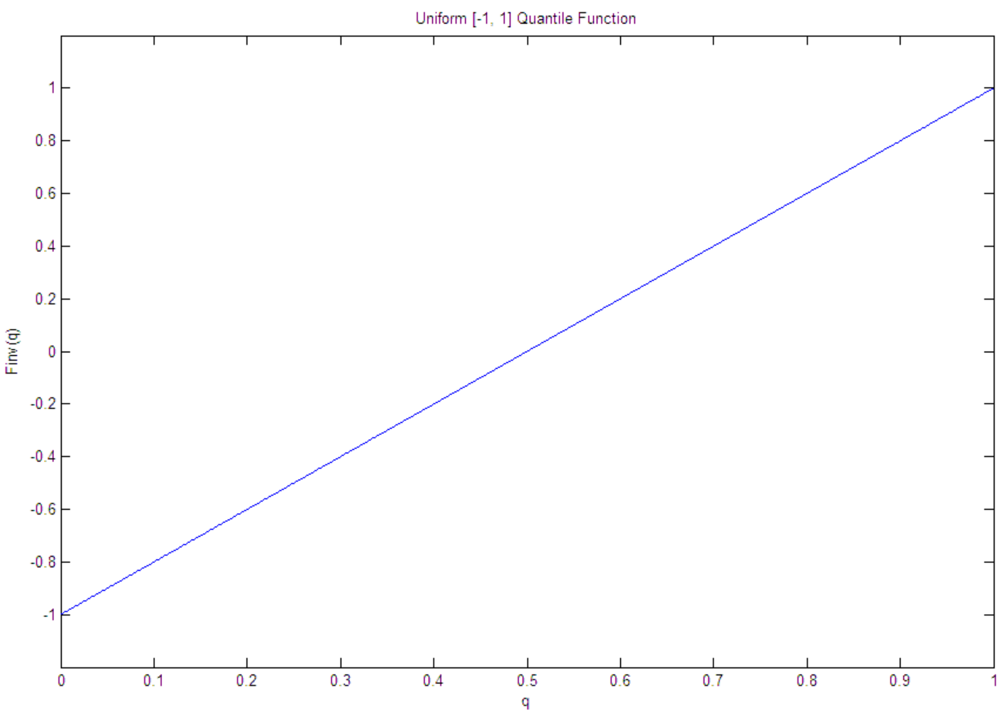
$$F(x) = \frac{x - a}{b - a}.$$

Then for $q \in (0, 1]$,

$$F^{[-1]}(q) = a + (b-a)q.$$

(a) Putting $a = -1$ and $b = 1$, the Uniform$(-1, 1)$ quantile function is

$$F^{[-1]}(q) = 2q - 1.$$



(b) Median $= F^{[-1]}(\tfrac{1}{2}) = 2(\tfrac{1}{2}) - 1 = 0$.

1st quartile $= F^{[-1]}(\tfrac{1}{4}) = 2(\tfrac{1}{4}) - 1 = -\tfrac{1}{2}$.

3rd quartile $= F^{[-1]}(\tfrac{3}{4}) = 2(\tfrac{3}{4}) - 1 = \tfrac{1}{2}$.

3 From week 1 exercise 6(b),

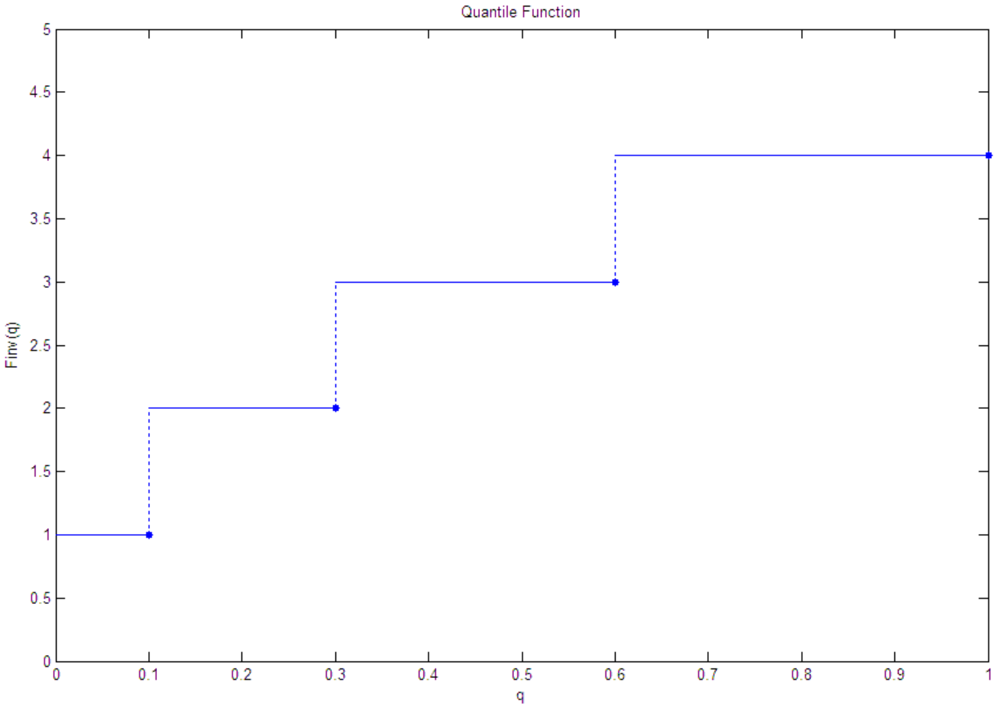
$$F(x) = \begin{cases} 0, & x < 1, \\ 0.1, & 1 \le x < 2, \\ 0.3, & 2 \le x < 3, \\ 0.6, & 3 \le x < 4, \\ 1, & 4 \le x. \end{cases}$$

Note that for a discrete CDF, the equality is to the left of x. If you look at the plot of the CDF, it is continuous from the right (or right-continuous) at $x = 1, 2, 3, 4$.

(a) The quantile function is, for $q \in (0,1]$,

$$F^{[-1]}(q) \begin{cases} 1, & 0 < q \le 0.1, \\ 2, & 0.1 < q \le 0.3, \\ 3, & 0.3 < q \le 0.6, \\ 4, & 0.6 < q \le 1. \end{cases}$$

The quantile function has the reversed property. The equality is now to the right of q, which means that in the plot of the quantile function, we will see that it is continuous from the left (or left continuous).



(b) Median $= F^{[-1]}(0.5) = 3$,

1st quartile $= F^{[-1]}(0.25) = 2$,

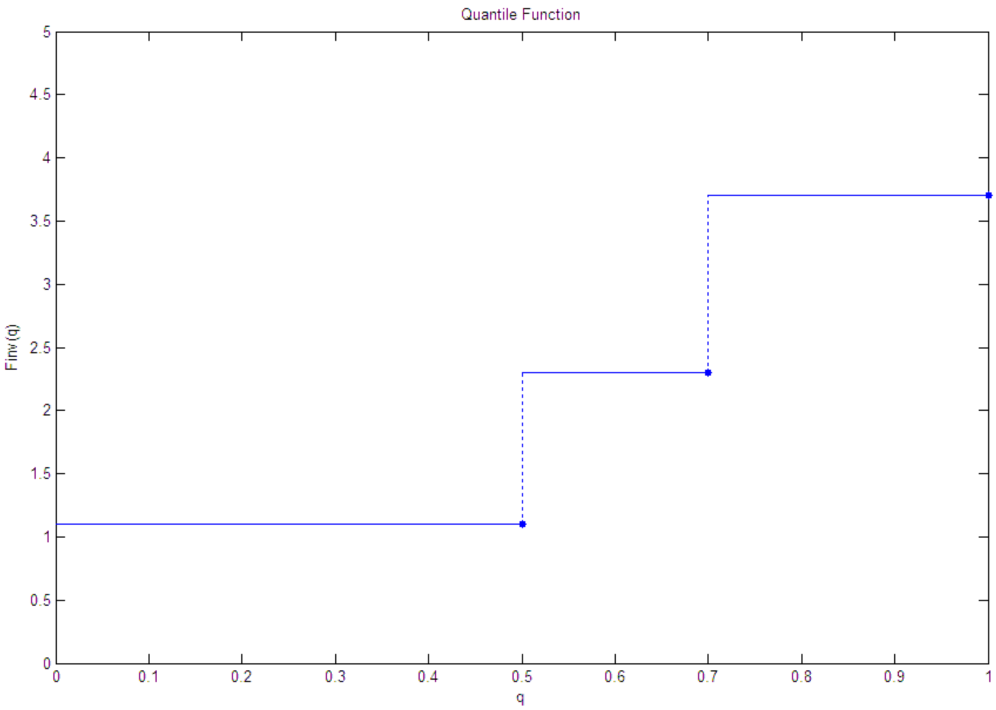
3rd quartile $= F^{[-1]}(0.75) = 4$.

④ From week 1 exercise 7,

$$F(x) = \begin{cases} 0 & , \quad x < 1\cdot 1, \\ 0\cdot 5 & , \quad 1\cdot 1 \le x < 2\cdot 3, \\ 0\cdot 7 & , \quad 2\cdot 3 \le x < 3\cdot 7, \\ 1 & , \quad 3\cdot 7 \le x . \end{cases}$$

(a) Quantile function is, for $q \in (0,1]$,

$$F^{[-1]}(q) \begin{cases} 1\cdot 1 & , \quad 0 < q \le 0\cdot 5, \\ 2\cdot 3 & , \quad 0\cdot 5 < q \le 0\cdot 7, \\ 3\cdot 7 & , \quad 0\cdot 7 < q \le 1 . \end{cases}$$



(b) Median $= F^{[-1]}(0\cdot 5) = 1\cdot 1$,

1st quartile $= F^{[-1]}(0\cdot 25) = 1\cdot 1$,

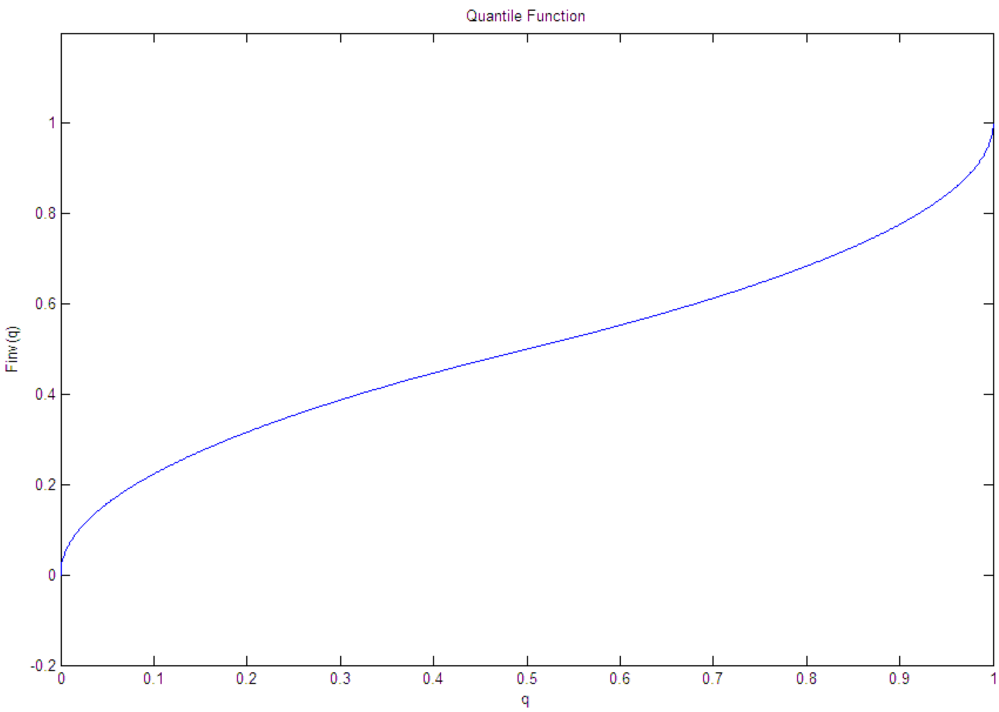
3rd quartile $= F^{[-1]}(0\cdot 75) = 3\cdot 7$.

(5) From week 1 exercise 8(b), for $x \in [0,1]$,

$$F(x) = \begin{cases} 2x^2 , & 0 \leq x < 0.5 , \\ -2x^2 + 4x - 1 , & 0.5 \leq x < 1 . \end{cases}$$

(a) Since the CDF is continuous, the quantile function is the pointwise inverse :

$$F^{[-1]}(q) = \begin{cases} \sqrt{q/2} , & 0 < q \leq 0.5 , \\ 1 - \sqrt{1 - \left(\frac{q+1}{2}\right)} , & 0.5 < q \leq 1 . \end{cases}$$

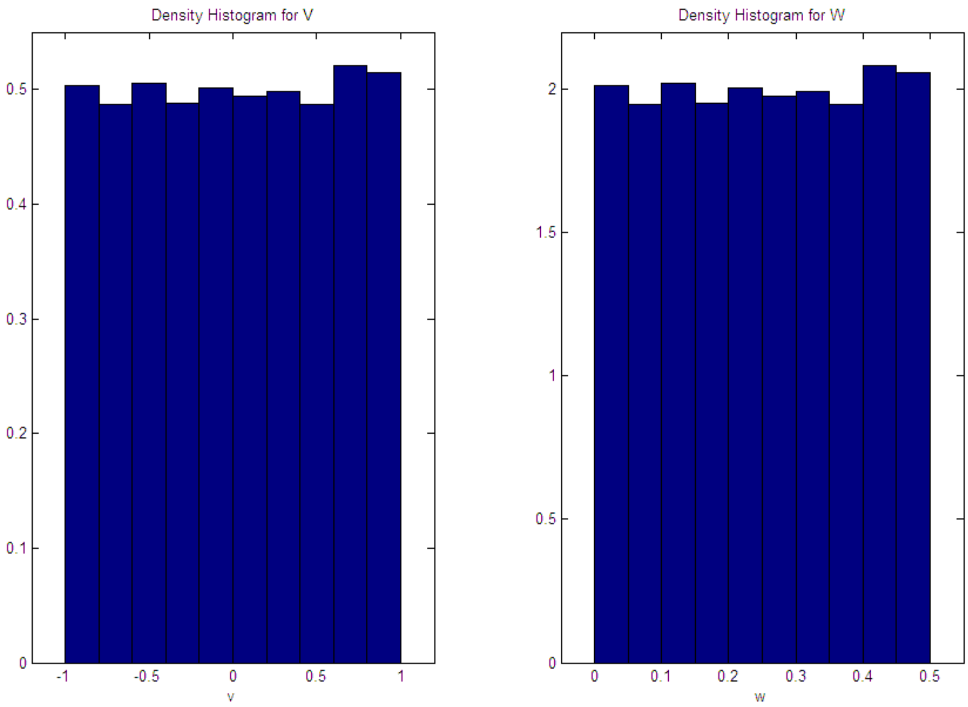


(b) Median $= F^{[-1]}(0.5) = \sqrt{1/4} = 0.5$,

1st quartile $= F^{[-1]}(0.25) = \sqrt{1/8} \simeq 0.3536$,

3rd quartile $= F^{[-1]}(0.75) = 1 - \sqrt{1/8} \simeq 0.6464$.

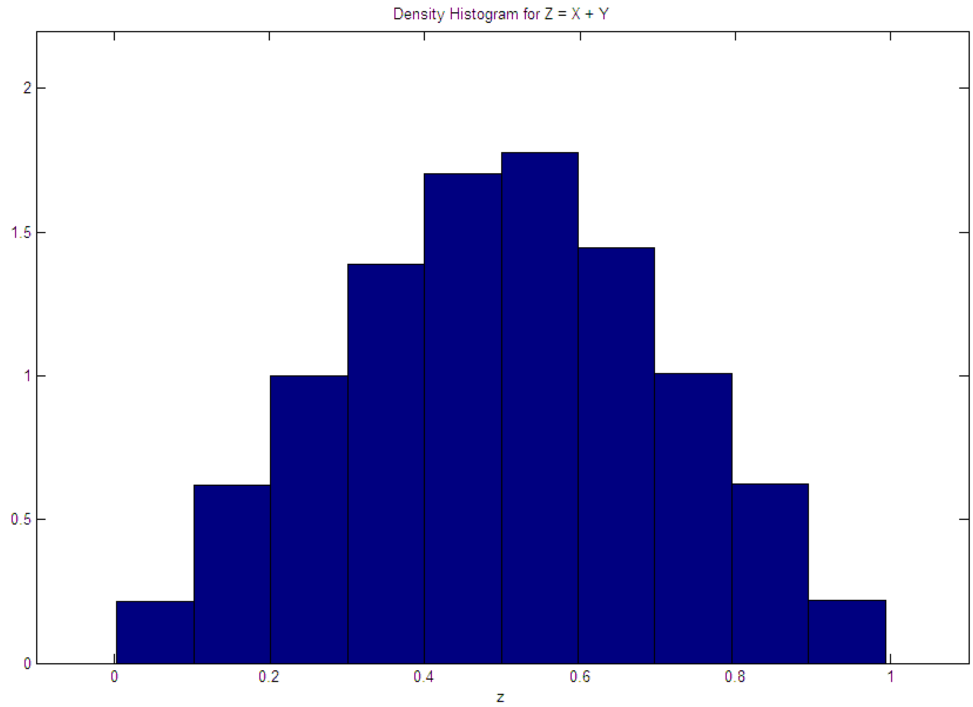
(6) (a)



(b) Let $X, Y \sim \text{Uniform}(0, \frac{1}{2})$ and let $Z = X + Y$.

(i) Support of X and Y is $[0, \frac{1}{2}]$, and so support of Z is $[0, 1]$.

(iii)



The shape of the PDF of Z is definitely not uniform but resembles the triangle density in week 1 exercise 8. In fact, it can be shown mathematically that the PDF of Z is the one given in exercise 8.