Exercise Set 12 - Solution

1 Exam-style Python questions [basic-normal]

a) `3`
Remember that the *first* item in a numpy (and generally Python) array has the index 0. So index 1 will actually give the second item in the sorted list.

b) `[13 12 17 15]` When performing a mathematical operation on a numpy array, it is performed on every item separately.

c) `(4,)`
`(2, 2)`
`myArray` is a a 1-d array (with length 4), whereas `reshapedArray` is a 2x2 array.

d) It will give an error message ("ValueError: operands could not be broadcast together with shapes (4,) (5,)"). As adding two numpy arrays (say a and b) means adding the first element of a to the first element of b etc., the two arrays have to be the same length.

e) `myUnbiasedStd = np.std(myArray,ddof=1)` makes sure that we devide by $N-1$.

f) The most straightforward (and fastest) way is to use the built-in minimum-finding function `np.min(myArray)`. For this particular code, as we have already sorted the numbers, we could also just look up the first number in the sorted array, using `sortedArray[0]`

g) - When adding up two lists, the second list is appended to the first list, whereas when adding two arrays, the values are added element by element.
- Multiplying a list $*2$ gives a list that is twice as long (for example `[1,2]*2` gives `[1,2,1,2]` but pultiplying a numpy array $*2$ multipiles each element times 2 and keeps the array size the same.
- Whereas numpy functions can be run on lists (they will be converted to arrays in the process), using attributes (such as `myArray.mean()` does no work for lists. Numpy arrays can be used in built-in mathematical functions (e.g. it is possible to write `myArray**2`), giving piece-wise operations. This does not work for lists. Note that something like `np.sin(myArray**2)` will also work on a list, as it gets converted to an array in the process.
- Generally (operations on) numpy arrays are also (faster) more memory efficient.

2 Proof of the variance of a sum [normal]

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}\left[((X+Y)-\mathrm{E}(X+Y))^2\right] \\
&= \mathrm{E}\left[([X-\mathrm{E}(X)]+[Y-\mathrm{E}(Y)])^2\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2+(Y-\mathrm{E}(Y))^2+2\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right]+\mathrm{E}\left[(Y-\mathrm{E}(Y))^2\right]+\mathrm{E}\left[2\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{E}\left[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)
\end{aligned}
$$

Alternatively you can start from $\mathrm{Var}(X+Y) = E[(X+Y)^2] - (E[X+Y])^2$

3 Skiing and Fondue [normal]

We run a 2-factor,2-level ANOVA. We denote skiing as the first and fondue as the second factor, and assign the 2 levels "0" and "1" to them. So, $X_{0,1,5}$ is the 5th person in the group that did not ski but had fondue. This person rated the day with 4, thus $X_{0,1,5} = 4$.

a) First we compute the group means:

$$\bar{X}_{0,0,\bullet} = 4.22, \quad \bar{X}_{0,1,\bullet} = 6.22, \quad \bar{X}_{1,0,\bullet} = 6.33, \quad \bar{X}_{1,1,\bullet} = 7.67$$

The partial (also known as marginal) means and the total mean are

$$\bar{X}_{0,\bullet,\bullet} = 5.22, \quad \bar{X}_{1,\bullet,\bullet} = 7.00, \quad \bar{X}_{\bullet,0,\bullet} = 5.28, \quad \bar{X}_{\bullet,1,\bullet} = 6.94, \quad \bar{X}_{\bullet,\bullet,\bullet} = \bar{X}_T = 6.11$$

We get a table of means:

.	no Skiing	Skiing	all
no Fondue	4.22	6.33	5.28
Fondue	6.22	7.67	6.94
all	5.22	7.00	6.11

b) Next we can calculate the sum squared errors within each group:

$$SS_{0,0} = 31.6, \quad SS_{0,1} = 25.6, \quad SS_{1,0} = 50.0, \quad SS_{1,1} = 20.0$$

This leads us to the sum of squares as:

$$SS_E = \sum_{i=1}^{2}\sum_{j=1}^{2} SS_{i,j} = 127.11$$

Note that if we had not been given the raw data, but only the unbiased estimator for the variance for each group and the number of elements in each group ($N_{Si,j}$) we could have used:

$$SS_E = \sum_{i=1}^{2}\sum_{j=1}^{2}(N_{Si,j} - 1)s_{i,j}^2$$

For the total SS, we need to go back to the raw data (the reference point is the total mean). We find:

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{9}(X_{i,j,k} - \bar{X}_T)^2 = 181.56$$

Looking at effects of each factor, and knowing that $I = J = 2$ and $N_S = 9$ for each group, we find:

$$SS_{B,skiing} = 9 * 2 * ((\bar{X}_{0,\bullet,\bullet} - \bar{X}_{\bullet,\bullet,\bullet})^2 + (\bar{X}_{1,\bullet,\bullet} - \bar{X}_{\bullet,\bullet,\bullet})^2) = 28.44$$
$$SS_{B,fondue} = 9 * 2 * ((\bar{X}_{\bullet,0,\bullet} - \bar{X}_{\bullet,\bullet,\bullet})^2 + (\bar{X}_{\bullet,1,\bullet} - \bar{X}_{\bullet,\bullet,\bullet})^2) = 25$$

This leads us to the ANOVA table below.

Source	ν	SS	MS	F
Skiing	1	28.44	28.44	7.39
Fondue	1	25	25	6.49
Error	33	127.11	3.85	
Total	35	181.56		

Using $\alpha = 0.05$ we find a critical $qF_{1,33} = 4.12$. Both factors have the same degrees of freedom, so they have the same critical F-value. Both factors well exceed the critical F-value, so both have a statistically significant effect on peoples happiness.

c) will be dealt with in the next exercise.