Problem Set 3: RVs and MATLAB exercises

Worth ~7.5 points (18/2.4) + Bonus 1.67 points (4/2.4)

Administrative comments:

a. Due date: Feb 10 (Mon), 11.59 pm. Remember: 5% off for each late hour!

(The remaining administrative and general comments are all the same as for PS2.)

b. This PS is meant to get you into the very trenches of data visualization in MATLAB. Rather than trying to set aside a large chunk of time on a single day to work on this problem set, splitting up work over a couple of days can help. Also, consider grappling with problem #2 last.

c. Upload your answer sheet and code to Gradescope. Note that there will be two assignment pages for each problem set, one for the PDF answer sheet and one for MATLAB code (single .m file, appropriately commented)
   a. Name your MATLAB code file as: ps#_name.m (e.g., ps1_shreesh.m)
   b. Within the code.m file, separate the code for each problem with a "section break", which, in MATLAB, is obtained by inserting a line with %% followed by the problem number. (e.g., %% Problem 2)
   c. Name all your figures as: ps#_name_figX.pdf (ps1_shreesh_fig3.pdf, for problem 3).

d. In general, efficiency of the code will be valued, as will be validity in the choice of variable names (per Tutorial under MATLAB-REVIEW module) and the clarity of the comments. Note: using the % symbol before a line of text comments out that line.

e. Collaboration/discussion are permitted (encouraged, in fact). However, make sure that the solutions you turn in are your own and that you understand everything that you put in your solutions. The goal, here, is to have a solid foundation upon which we can build in the coming weeks.

f. Always abide by JHU honor code (see course syllabus).

General tips and comments: For all the problems,

i. Label the x,y (& z) axes appropriately. [Hint: "xlabel" (and ylabel and zlabel)]

ii. Make sure that the axis labels are of fontsize 15 [Hint: h=xlabel("blood pressure"); set(h,"fontsize",…)]

iii. Set the font size of the x and y tick labels to be 10. ["help set"; the syntax that you will ultimately use will look like: set(gca,"fontsize",…)].

iv. Provide a title for all plots [hint: "help title"]. Set title's fontsize to be 18.

v. Use the command "print –dpdf figName" to save the current figure as a pdf file. The "current" figure is the one that you most recently clicked on with your mouse.

Qn	Points	Should take you (in min)	How long did it take? (approximately, min)
1	**2**	**15'**	**10'**
2	**2**	**15'**	**20'**
3	**5**	**30'**	**60**
4	**5**	**15'**	**30'**
5 (BONUS)	**5**	**?**	**45'**

*Copy and paste this table into your answer sheet,
(or) just use this word document to type out your written answers and then rename it as described in "c" above

---

1. Random numbers (2 point). Generate a vector with 30 random numbers drawn from a normal distribution with a mean of 3.2 and a variance of 16. (Hint: help randn)
This is done in the problem set code.

JHU.PBS. AS.200.659

2. Variance property (2 points). Var(X) = E[$(X-\mu_x)^2$]. Derive what Var(aX) is. (Hint: Replace X with aX on the both sides of the formula and expand things out)

As the hint states, we can first replace X with aX on both sides of our formula:

$$Var(aX) = E(aX - u_x)^2$$

Next, we know that the value of u_x is the same as $E[X]$, or the *mean* of our random variable X. Substituting this, we have:

$$Var(aX) = E[(aX - E[aX])^2]$$

We can use the linearity of expectation, which states that any constants a and b for any two random variables X and Y:

$$E[aX + bY] = aE[X] + bE[Y]$$

Effectively, the expected values of the variables can be "pulled out" and multiply the expected value. Thus, we now have:

$$Var(aX) = E[(aX - aE[X])^2]$$

We can now resubstitute u_x for $E[X]$:

$$Var(aX) = E[(aX - au_x)^2]$$

Next, factor out the a from the expression:

$$Var(aX) = E\left[\left(a(X - \mu_x)\right)^2\right]$$

Performing this step gives us:

$$Var(aX) = E[a^2(X - \mu_x)^2]$$

Using the linearity of expectation again, we can "pull out" the constant a^2:

$$Var(aX) = a^2E[(X - \mu_x)^2]$$

We know that $E[(X - \mu_x)^2]$ is the variance of our random variable! So we ultimately have:

$$Var(aX) = a^2Var(X)$$

This demonstrates a key property of how a constant applied to a random variable can either stretch or shrink the spread of a random variable! Deviations from the mean are squared when the variable is multiplied by a constant!

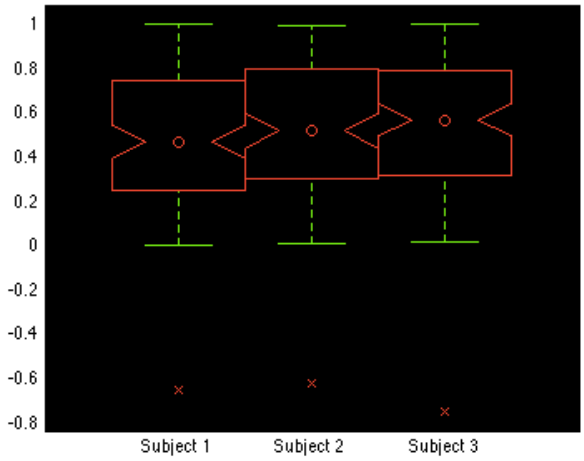
3. Boxplot (5 points). Load the data set and produce a boxplot of the data to have it match (mostly) the shown figure.

a. (1) Set color of boxes to red (don't worry about turning some of the lines green, as in the adjacent plot).

This is shown in figure ps3_3a

b. (1) Plot the median as a dot instead of as a line.

This is shown in ps3_3b. I could not achieve showing this as a dot in the style shown here by changing the MedianStyle to "target" as any customization I applied would invariably yield a red circle with a white center. Strangely, selecting "MedianInner" and "MedianOuter" for customization yielded no influence on the plot. I found these properties of the graph through the get method. The way I did it feels inefficient, but I'm not sure how else it is possible.



c. (1) Change the outlier symbol to 'x'

This is shown in ps3_3c.

d. (1) Set the xticklabels to be 'Subject 1', 'Subject 2', and 'Subject 3'.

This is shown in ps3_3d. Figure ps3_3d will be the complete figure for this problem with all formatting included.

e. (1) What does the interquartile range of a distribution mean (briefly and intuitively)? Show how you would calculate it for a dataset (say a vector x, containing 10 reaction times)?

For question e., write your answer down as comments below the code for a.-f (properly sectioned). Answer in equation(s) and sentence(s).

This is shown in the code of the problem set.

Hint: Everything you need for this, you should find under "help boxplot" + Slides_MATLAB_basics_part2.ppt.
Dataset**: ps3_3_boxplot.mat**

4 (5 points) Given two random variables X and Y, what are the mathematical/statistical definitions for:

a. Uncorrelatedness

When describing the **uncorrelatedness** between the random variables X and Y, we must first describe the **covariance** between them.

The covariance is mathematically defined as:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Further, it is simplified as described in the class's optional reading into:

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

In words, this means that the covariance is a calculation of the expected value of the product of deviations of X and Y from their respective means. Expected values simply describe the average outcome you would expect to see if you sampled your variables many times. It is critical to notice that while the covariance can tell you the directionality of a relationship (i.e. both increase together or both decrease together), it does not tell you how steep the slope is that describes this linear relationship or how far these variables deviate from the linear relationship described.

Now that we understand covariance, we can establish what it means for something to be uncorrelated. As described in the reading, two random variables X and Y are **uncorrelated** if and only if the covariance between them is zero! Represented mathematically, we see the following:

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0$$

Or:

$$E(XY) = E(X)\,E(Y)$$

In words, this means that the average of the products of the values of X and Y is equal to the product of the average of X values and the average of Y values. Thus, there is no linear relationship between the values of X and Y!

b. Independence

As described in the optional reading, two events or random variables are independent if and only if their joint probability equals the product of their probabilities. In other words, if two variables are independent there is no relationship between them of any kind! Written mathematically, you see the following:

$$P(X \cap Y) = P(X)\,P(Y)$$

A joint probability is the probability of two or more events happening together and thus captures the likelihood of the co-occurrence of our two variables. If the two variables are independent, the occurrence of one event does not influence the other.

c. Orthogonality

Our reading demonstrates that two random variables, X and Y, are orthogonal if and only if the expected value of the product of events X and Y equal zero! In other words, it is the average value of the product of our two variables. Mathematically, this is simply represented as:

$$E(XY) = 0$$

d. Uncorrelated and zero mean RVs => independent. T/F? Explain briefly.

This statement is false! Random Variables that are uncorrelated with zero mean does not imply independence for our random variables. If the variables are uncorrelated – meaning, there is no linear relationship between them – it is still possible that there are non-linear forms of dependence between them. A zero mean value does not describe independence either but rather only means that the expected value of the variables is zero.

e. Uncorrelated => orthogonal. T/F? Explain briefly.

In the general case, this statement is false! While it is true that the two random variables will be uncorrelated if their covariance is zero and there is therefore no linear relationship between them, there is still the chance that there is still a non-linear relationship between them. Further, the two variables are orthogonal if the expected value of the product of X and Y is zero. The only case where uncorrelated variables are orthogonal is where both variables X and Y have zero means! Mathematically, we would see:

$$E(XY) = E\big[(X - E[X])\big(Y - E(Y)\big)\big]$$

5. BONUS **(4 points)** Why is the sample variance calculated by dividing by (n-1) rather than n? Please show this explicitly with a derivation. (Hint: The answer to this was fully worked out at the end of Lecture 03A_slides. I would like you to not only write it out, but also explain how / why you are able to go from one step to another. Either hand-written and attached as a picture in the same doc, or typed are fine.)

First, we know that the population variance can be defined as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

Where N is the total number of values in a population, μ is the true population mean, and $\sigma^2$is the true population variance. Since we do not know the true mean or number of values in our population, however, we must estimate the sample variance which is defined as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2$$

Where n is the number of samples,$\overline{X}$ is the sample mean, and s^2 is the sample variance.

To derive why we use $n - 1$, it is useful to take the expectation of our equation here. We need to demonstrate that $E(s^2) = \sigma^2$and so we take the expected value of each component of our equation. That looks like the following:

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

Now, we can "pull out" the $\frac{1}{n}\sum_{i=1}^{n}$ term of our equation as it is a constant that can be removed due to the linearity of expectation and perform the expansion of our square term:

$$E[\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}E\left[y_i^2 - \frac{2}{n}y_{i\sum_{j=1}^{n}y_j} + \frac{1}{n^2}\sum_{j=1}^{n}y_j\sum_{j=1}^{n}y_k\right]$$

Next, we can get the expected value of each term and pull out all constants from our equation again using the linearity of expectation being distributed across our terms:

$$E[\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E\big[y_i^2\big] - \frac{2}{n}\sum_{j=1}^{n}E\big[y_iy_j\big] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}E\big[y_j\,y_k\big] + \frac{1}{n^2}\sum_{j=1}^{n}E\big[y_j^2\big]\right]$$

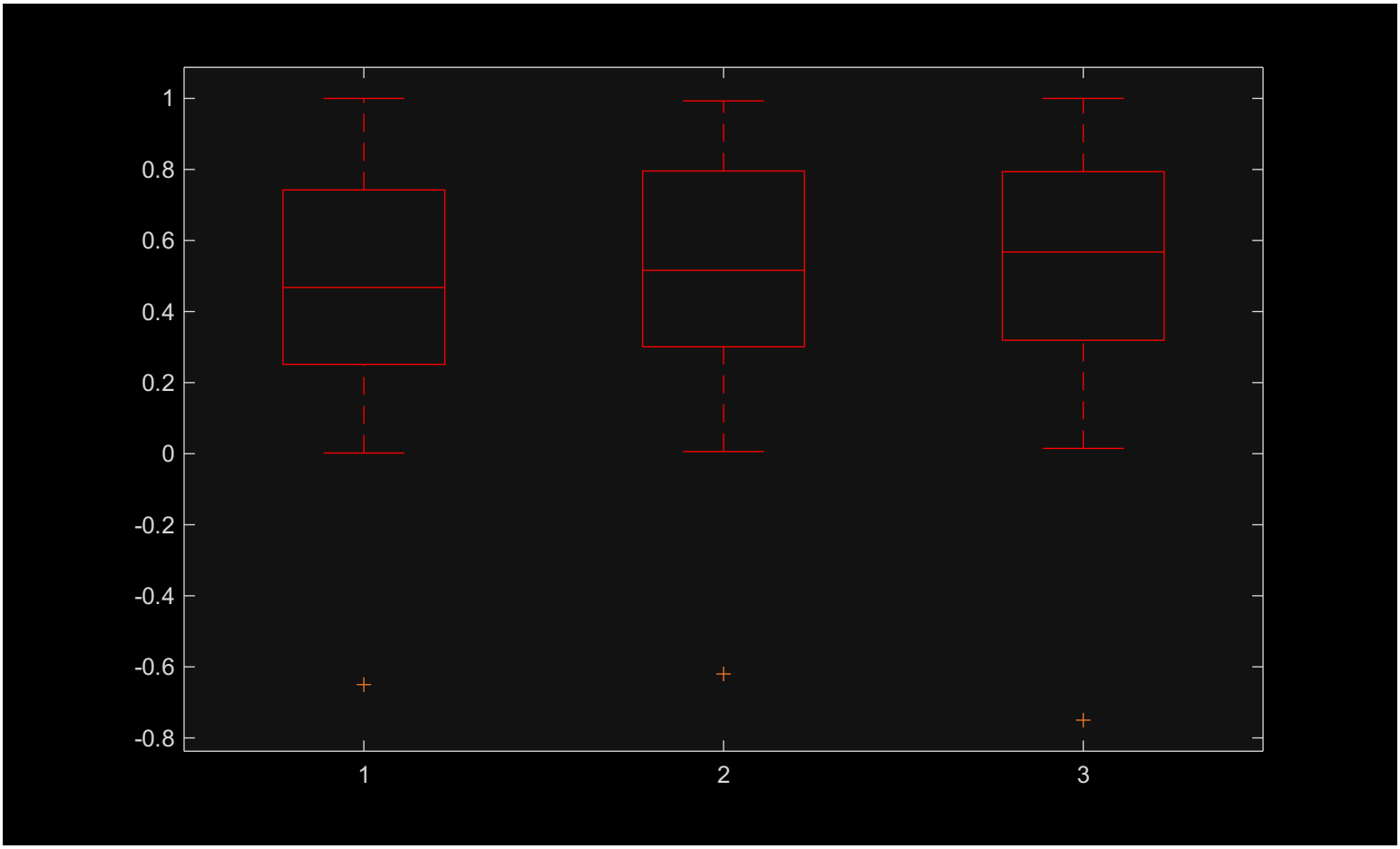
Next, we know that that each value sampled from our random variable is independent and, over a large number of samples, the expected value is simply equal to the mean μ of our variables! So $E\big[y_j\big] = E[y_k] = \mu$. Substituting this into our equation above gives us:

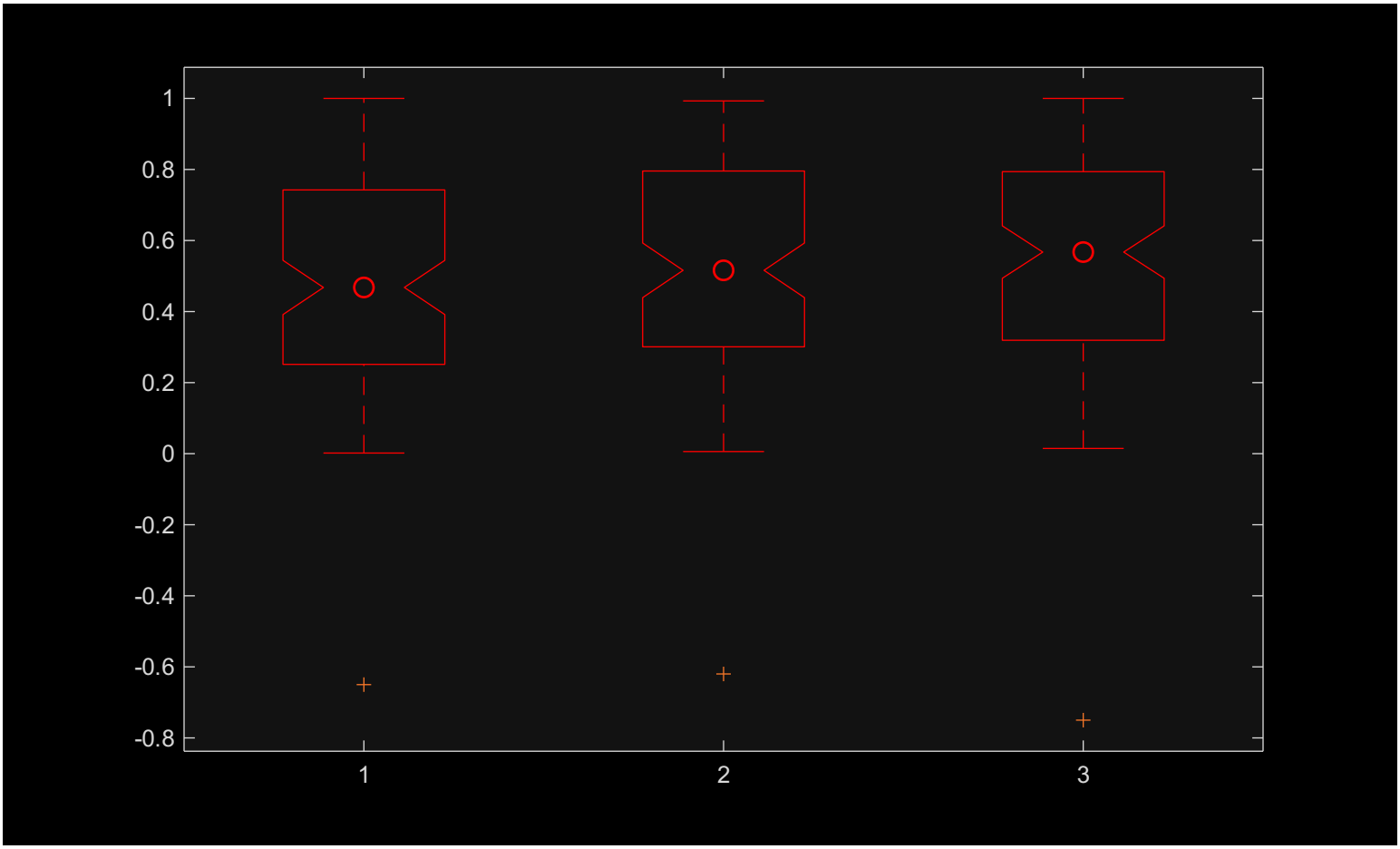
$$E[\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2 + \mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2 + \mu^2)\right]$$

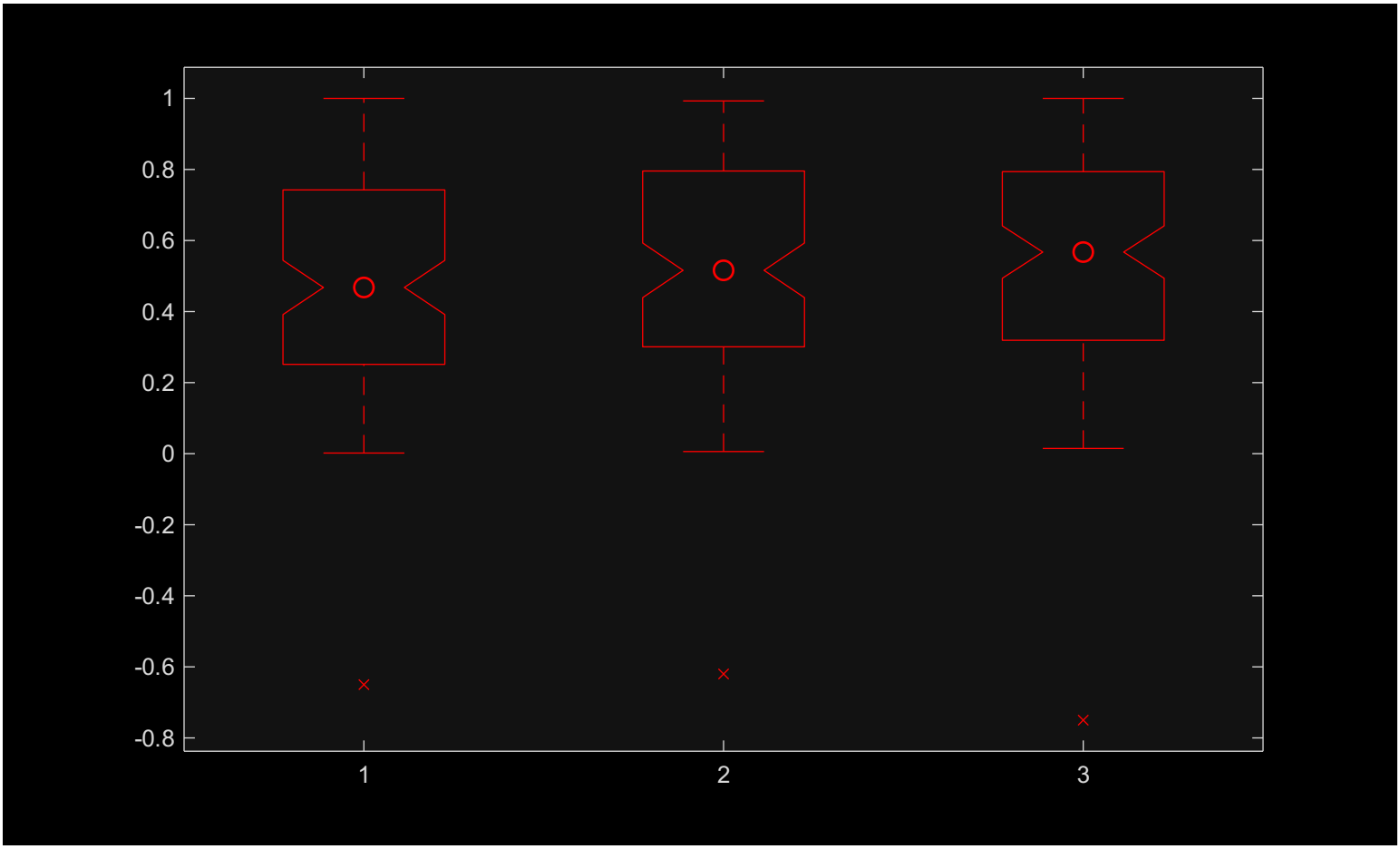
If we simplify this, we yield:

$$E[\sigma^2] = \frac{n-1}{n}\sigma^2$$

We can see that our variance is "biased" by a constant factor of $\frac{n-1}{n}$here! We can use Bessel's correction and multiply by the inverse of the constant to ultimately yield an unbiased estimator of just σ^2!


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1
2
3


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1
2
3


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1
2
3


Subject Data Boxplots
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Data
Subject 1
Subject 2
Subject 3
Subjects