Lecture 10 Worksheet

Chrysafis Vogiatzis

Every worksheet will work as follows.

1. You will be entered into a Zoom breakout session with other students in the class.
2. Read through the worksheet, discussing any questions with the other participants in your breakout session.
   - You can call me using the "Ask for help" button.
   - Keep in mind that I will be going through all rooms during the session so it might take me a while to get to you.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students in a breakout session is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see Compass) to submit your work.

Worksheet 1: Normally distributed random variables

We begin this section by mentioning something very important:

Adding two (or more) independent normal distributed random variables together results in a normal distribution. In mathematical terms: if $X_i, i = 1, \ldots, n$ are independent random variables distributed normally, then $Y = \sum_{i=1}^{n} X_i$ is also normally distributed.

To fully describe a normally distributed random variable, we need its expectation and variance.

Problem 1: Expectation and variance of the sum of normally distributed random variables

Consider two normally distributed random variables X and Y, which are independent. What is the expectation and variance of $X + Y$?

Answer to Problem 1.

In general, $Y = \sum_{i=1}^{n} X_i$ is normally distributed with $\mu = E[Y] = \sum_{i=1}^{n} E[X_i]$ and $\sigma^2 = Var[Y] = \sum_{i=1}^{n} Var[X_i]$.

Problem 2: Application

You own a portfolio of stocks that consists of 5 stocks S_1 and 10 stocks S_2. Let X_1 be the price of stock S_1 and X_2 the price of stock S_2 one year from now. X_1 is normally distributed with $\mathcal{N}(15, 100)$ and X_2 is normally distributed with $\mathcal{N}(20, 100)$. Last, assume that the two stocks are totally unrelated.

What is the probability that your portfolio is worth more than $300 one year from now? What is the probability that your portfolio is worth less than $200 one year from now?

Answer to Problem 2.

Problem 3: Comparing random variables

Consider two normally distributed random variables X and Y, which are independent. What is the expectation and variance of $X - Y$? Can you make the claim that $X - Y$ is normally distributed? [1]

Answer to Problem 3.

[1] Can you say that $Z = -Y$ is normally distributed? If so, what can you say about $X + Z$?

Problem 4: Comparing stock prices

Again, consider that X_1 is the price of stock S_1 and X_2 is the price of stock S_2 one year from now. Both are normally distributed: $X_1 \sim \mathcal{N}(15, 100)$ and $X_2 \sim \mathcal{N}(20, 100)$. Finally, asume that X_1, X_2 are independent.

What is the probability that S_1 is worth more than S_2 a year from now? [2]

[2] Think of $Z = X_1 - X_2$...

Answer to Problem 4.

What about the case where two random variables (distributed normally) are **not independent**? One could make the argument that two stocks affect each other, and knowing that one has gone up may affect our perspective of the other one going up, too. In this case, where independence is hard to assume and use, can we still compare two stocks (or calculate the probability of our portfolio being above a threshold)?

The answer is yes! However, we need to introduce ideas like **dependence** and measure it with **covariance**. More on that in Lecture 12...

Worksheet 2: Using the central limit theorem

In the online course world, certain classes have thousands of students: these classes are sometimes referred to as *Massive Open Online Courses* (MOOC). On average there are 8000 students in a course. For the purposes of this exercise, we assume that a student successfully finishes a MOOC 7.5% of the time and that all students behave independently. Assume that probability is the same regardless of the course. Finally, assume that every class has exactly 8000 students.

Answer the following questions.

Problem 5: Setting up the distribution

What is the best distribution to model the number of students to successfully finish the class? What is the mean and the variance of the number of students who successfully finish the class? [3]

[3] Recall: you have 8000 students, each of whom may succeed or fail..

Answer to Problem 5.

Problem 6: Setting up the central limit theorem

Now, consider the case of an online course provider, such as Coursera. Assume the total number of classes the provider offers is equal to 1000. For each individual class, the number of students who successfully finish that specific class follows the distribution you found in Problem 5.

What distribution does the average number of students graduating from all class of the provider follow? The average number of students graduating in all 1000 classes can be found by summing the number of graduates in each class and dividing by 1000. [4]

[4] Does the central limit theorem apply? If so, then the distribution is...

Answer to Problem 6.

Worksheet 3: Setting up hypothesis testing

Recall that we have made the hypothesis that a student successfully finishes a class 7.5% of the time. We have also found that the number of students successfully finishing one class follows a binomial distribution (based on Problem 5) and the average number of students across 1000 courses follows a normal distribution (based on Problem 6).

Problem 7: Checking the numbers

The online course provider has decided to check whether this "7.5%" success rate is true or not. They have decided to survey 15 of the 1000 courses they offer and they found the following: [5]

Course 1	638	Course 2	695	Course 3	682
Course 4	673	Course 5	685	Course 6	691
Course 7	694	Course 8	661	Course 9	680
Course 10	678	Course 11	619	Course 12	687
Course 13	685	Course 14	727	Course 15	660

What should the average number of graduating students in these 15 classes be distributed as? Once again, to find the average number of graduates, simply add up the numbers for the 15 classes and divide by $n = 15$. [6] What is the mean and the variance of the distribution?

Answer to Problem 7.

Problem 8: Using the central limit theorem

From the numbers they found (see the tabulated data from Problem 7) that on average a whooping 677 students in each class successfully finishes it.

Based on the distribution you have identified in Problem 7, what is the probability that there are more than 650 students on average finishing each class? [7]

[5] You will not need the numbers until Problem 8.

[6] Treat 15 as a large enough number for the central limit theorem to apply.

[7] Remember that z values that do not appear in the table (i.e., are larger than 3.9) can be treated as corresponding to 1 (100%)!

Answer to Problem 8.

Problem 9: Huh?

Hopefully, you have gotten a very, very small [8] probability for Problem 8. This implies that the numbers they got appear to be highly improbable. While we are at it, what is the probability that the number of students that (on average) finish successfully each course is more than 610?

[8] Possibly zero!

Answer to Problem 9.

Problem 10: Rejecting a hypothesis

Considering the small probability you got for the average to be as high as 610 or more, what can you deduce for the success rate of 7.5%? Do you think it is valid? Or is the true success rate higher/lower? [9]

[9] Think: if the success rate is higher, and the mean of the normal distribution (for the average of 15 classes) is also higher, then does that mean the probabilities in Problem 8 and 9 go up or down?

Answer to Problem 10.

Congratulations, you just performed a fully-fledged data analysis experiment! We will focus a lot on this part after our second midterm.

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000