CIS 2033, Spring 2017[1]

Instructor: David Dobor

April 17, 2017

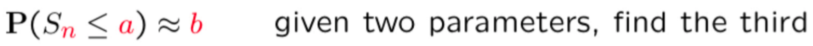
We now cover four essential examples that illustrate the types of questions that we usually answer using a normal approximation based on the central limit theorem.

Introduction

We will now go through a sequence of examples that illustrate the different types of questions that we usually answer using a normal approximation based on the central limit theorem. In general, the setup is as follows. Consider this event: the sum of n i.i.d. random variables, S_n, is less than or equal a certain number, a. We are interested in the probability of this event being approximately equal to some other number, b:

$$\mathbf{P}(S_n \leq a) \approx b$$

Notice that this statement involves three parameters, a, b, and n, and you can imagine problems where you are given two of these parameters, and you're asked to find the third.



$\mathbf{P}(S_n \leq a) \approx b$ given two parameters, find the third

This gives us the different variations of the questions that we might be able to answer. We will look at four such variations by going through examples.

The setting for the four representative problems.

The setting will be as follows. We are loading a container with packages. Each package has a random weight, X_i, which is a random variable that's drawn from an exponential distribution with a parameter $1/2$. Moreover, all of the X_is are independent.

[1] Supplemental reading: Bertsekas Textbook Chapter 7.

The Four Representative Examples

EXAMPLE 1. First, suppose that we load the container with 100 packages. We would like to calculate the probability that the total weight of the 100 packages exceeds 210 (which might be the capacity of the container).

- Load the container with $n = 100$ packages.
- Calculate $\mathbf{P}(S_n \geq 210)$.

Since we will be using the central limit theorem, we will have to work with the standardized version of S_n in which we subtract the mean of S_n and divide by the standard deviation of S_n.

Standardized S_n, denoted by Z_n.

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

Of course, to compute Z_n we will need to know the mean and the standard deviation of S_n. Now for an exponential, X_i, the mean is the inverse of lambda, $\mu = \mathbf{E}[X_i] = 1/\lambda = 2$, and the standard deviation is also the inverse of lambda, $\sigma = \sigma_{X_i} = 1/\lambda = 2$.

In our examples, we take $\mu = \sigma = 2$.

Then, the next step is to take the event $S_n \geq 210$ and rewrite it in a way that involves the random variable Z_n. We've done this before: We take the original description of the event, $S_n \geq 210$, subtract from both sides of the inequality the number $n\mu$, in this case $100 \cdot 2$, and divide the result by $\sqrt{n}\sigma = \sqrt{100} \cdot \sigma = 20$:

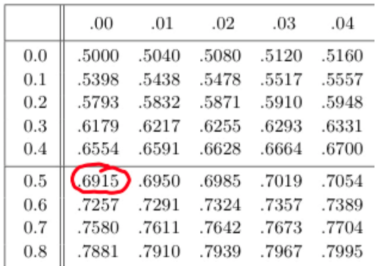
$$\begin{aligned}\mathbf{P}(S_n \geq 210) &= \mathbf{P}\Big(\frac{S_n - 200}{20} \geq \frac{210 - 200}{20} \Big) \\ &= \mathbf{P}(Z_n \geq 0.5)\end{aligned}$$

Why did we do this? Well, $Z_n \geq 0.5$ is just an equivalent representation of the original event, $S_n \geq 210$, and at this point, we can use the central limit theorem approximation to say that the probability $\mathbf{P}(Z_n \geq 0.5)$ is approximately the same if we use a standard normal instead of Z_n.

Z_n is approximately standard normal, by the Central Limit Theorem.

Now, for a standard normal we can calculate probabilities in terms of the CDF that's given in the table. Thus we finalize our calculations by looking up the appropriate entries in the normal table:

$$\begin{aligned}\mathbf{P}(S_n \geq 210) &= \mathbf{P}(Z_n \geq 0.5) \\ &\approx \mathbf{P}(Z \geq 0.5) = 1 - \mathbf{P}(Z < 0.5) \\ &= 1 - \Phi(0.5) = 1 - 0.6915 \\ &= 0.3085\end{aligned}$$



	.00	.01	.02	.03	.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389
0.7	.7580	.7611	.7642	.7673	.7704
0.8	.7881	.7910	.7939	.7967	.7995

Figure 1: A fragment of the standard normal table. The circled quantity is used in our calculations.

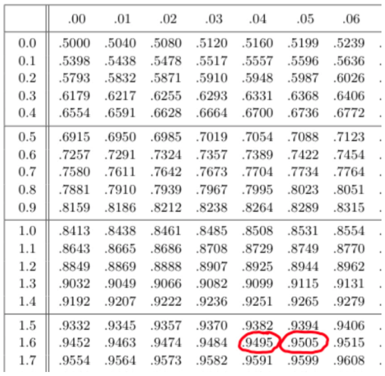
Example 2. In the next example, we ask a somewhat different question. We fix again the number of packages to be 100, but we are given some probabilistic tolerance. We allow the total weight of the packages to exceed the capacity of the container. But we don't want that to happen too often, we want to have only 5% probability of exceeding that capacity.

- Again, load $n = 100$ packages.
- Let package weights X_i be exponential with $\lambda = 1/2$.
- Goal: Choose the "capacity", a, so that $\mathbf{P}(S_n \geq a) \approx 0.05$.

How should we choose the capacity of the container if we want to have this kind of a specification?

So we proceed as follows. We want 0.05 to be approximately equal to $\mathbf{P}(S_n \geq a)$. So we take the event $S_n \geq a$ and rewrite it in terms of the standardized random variable, as in Example 1.

$$\begin{aligned} 0.05 &\approx \mathbf{P}\left(\frac{S_n - 200}{20} \geq \frac{a - 200}{20} \right) \\ &\approx \mathbf{P}\left(Z_n \geq \frac{a - 200}{20} \right) \\ &\approx 1 - \Phi\left(\frac{a - 200}{20} \right) \end{aligned}$$



	.00	.01	.02	.03	.04	.05	.06	
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.

Figure 2: A fragment of the standard normal table. The circled quantities are used in our calculations.

Now, what this tells us is that the quantity $\Phi\left(\frac{a-200}{20} \right)$ should be equal to $1 - 0.05 = 0.95$.

And what does this tell us about the event $\frac{a-200}{20}$ itself? We look at the table and try to find in it an entry of 0.95.

Well, we don't find the exact entry of 0.95, but we could choose either of the two quantities that are circled in Figure 2. Or we might decide to split the difference and say that we get the value of 0.95 when the argument is 1.645.

So we conclude that in order for $\Phi\left(\frac{a-200}{20} \right)$ to be 0.95 we need

$$\frac{a - 200}{20} = 1.645,$$

which we solve for a to find that

$$a = 232.9.$$

And this concludes our second example, showing how to choose the capacity of the container.

EXAMPLE 3. Our next example is a little more challenging. Here, we will fix a and b and we will ask for the value of n. Here's a type of question that has this flavor. We are given the capacity of our container. We want to have a small probability of exceeding that capacity. How many packages should you try to load?

- Again, let package weights X_i be exponential with $\lambda = 1/2$.
- Let the container capacity, a, be equal to 210.
- Let the probability of exceeding that capacity, b, be equal to 0.05.
- Question: How large can n be, so that $\mathbf{P}(S_n \geq 210) \approx 0.05$?

What is the value of n for which this relation will be true?

So we proceed, as usual, by taking this event and rewriting it in a way that involves the standardized version of S_n.

$$\mathbf{P}\Big(\frac{S_n - 2n}{2\sqrt{n}} \geq \frac{210 - 2n}{2\sqrt{n}} \Big)$$
$$\approx 1 - \Phi\Big(\frac{210 - 2n}{2\sqrt{n}} \Big)$$

Now we want the quantity $1 - \Phi\Big(\frac{210-2n}{2\sqrt{n}} \Big)$ to be approximately equal to 0.05, which, once more, means that

$$\Phi\Big(\frac{210 - 2n}{2\sqrt{n}} \Big) \approx 0.95$$

Looking back at the fragment of normal table given in Figure 2, we conclude, as in example 2, that we want to solve the following for n:

$$\frac{210 - 2n}{2\sqrt{n}} = 1.645$$

Here, we get an equation for n. Unfortunately, it is a quadratic equation, but we can solve it (even if it shows up on a quiz). And after you solve it, numerically or using the formula for the solution of quadratic equations, you find the value of n that's somewhere between 89 and 90.

Now, n is an integer, so you could choose either 89 or 90. If you want to be conservative, then you would set n to the smaller value of the two and conclude that $n = 89$.

EXAMPLE 4. Our last example is going to be a little different. Here's what happens. We start loading the container that has a capacity of 210. We keep loading the packages, but once we load a package and we see that the total weight of the packages loaded so far has exceeded 210, we stop.

Let N be the number of packages that have been loaded. This number is random: if you're unlucky and you happen to get lots of heavy packages, then you will stop earlier.

We would like to calculate, approximately, the probability that the number of packages that have been loaded is larger than 100.

- Again, let package weights X_i be exponential with $\lambda = 1/2$.
- Let the container capacity, a, be equal to 210.
- Load the container until the total package weight exceeds 210, and let N be the number of packages loaded. Note that N is a random variable.
- Question: What is $\mathbf{P}(N > 100)$, the probability that the number of packages that have been loaded is larger than 100?

Now, this problem feels a little different. The reason is that N is not the sum of independent random variables and so we do not have a version of the central limit theorem that we could apply to N. What can we do?

Here's how we go about it. We take the event $N > 100$ and express it in terms of the X_is. What does it mean that we loaded more than 100 packages? It means that at the time we were loading the 100th package, we didn't stop. And this means that at that time, after we loaded the 100th package, the weight had not exceeded 210. So the event that we're dealing with here is the same as the event that the first 100 packages have a total weight which is less than or equal to 210:

$$\mathbf{P}(N > 100) = \mathbf{P}\Big(\sum_{i=1}^{100} X_i \leq 210 \Big)$$

But now we're back to a problem that we know how to solve. The way to solve it is to take the random variable $\sum_{i=1}^{100} X_i$ and standardize it – this actually is essentially the same calculation as in our very first example – to get the following

$$\begin{aligned}\mathbf{P}(N > 100) &= \mathbf{P}\Big(\sum_{i=1}^{100} X_i \leq 210 \Big)\\ &\approx \Phi\Big(\frac{210-200}{20} \Big)\\ &= \Phi(0.5) = 0.6195\end{aligned}$$

And this concludes our last example.

These four examples that we worked through cover pretty much all of the types of problems that you might encounter. Of course, sometimes it might not be entirely obvious what kind of problem you are dealing with. You may have to do some translation from a problem statement to bring it in the form that we dealt with here. But once you bring it into a form where you can get close to applying the central limit theorem, then the steps are pretty much routine, as long as you carry them out in a systematic and organized manner.