4/14/2022

Chapter 7: The Central Limit Theorem

The Central Limit Theorem basically quantifies the idea behind the Law of Large Numbers.

As our sample size increases, the proportion (or means) will tend to get closer and close to the true value.

The Central Limit Theorem goes further. It allows us to calculate the probability of being a certain distance from the true value. It allows us to calculate a margin of error: how far off from the true value are we likely to be? Given a certain probability. Typically, that probability is 95%.

Two cases: means, and proportions.

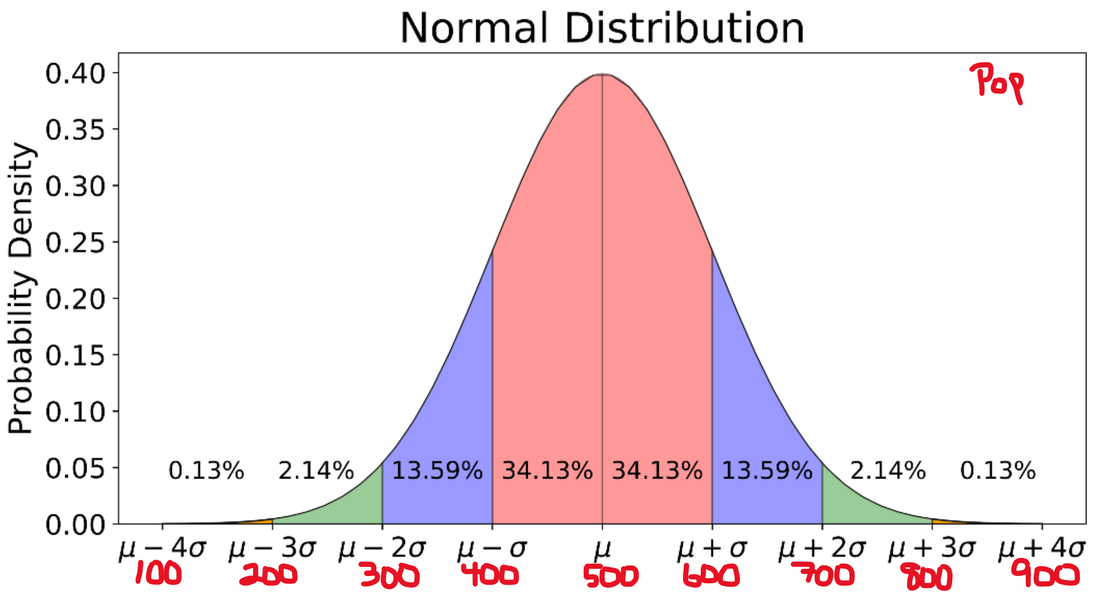
Means:
The central limit theorem states that sample mean $\bar{x}, \mu_X$ (of size n) tends to be distributed around the mean of the population μ (true mean), and the distribution has a standard deviation of $\frac{\sigma}{\sqrt{n}}$. (sigma is the standard deviation of the population.)

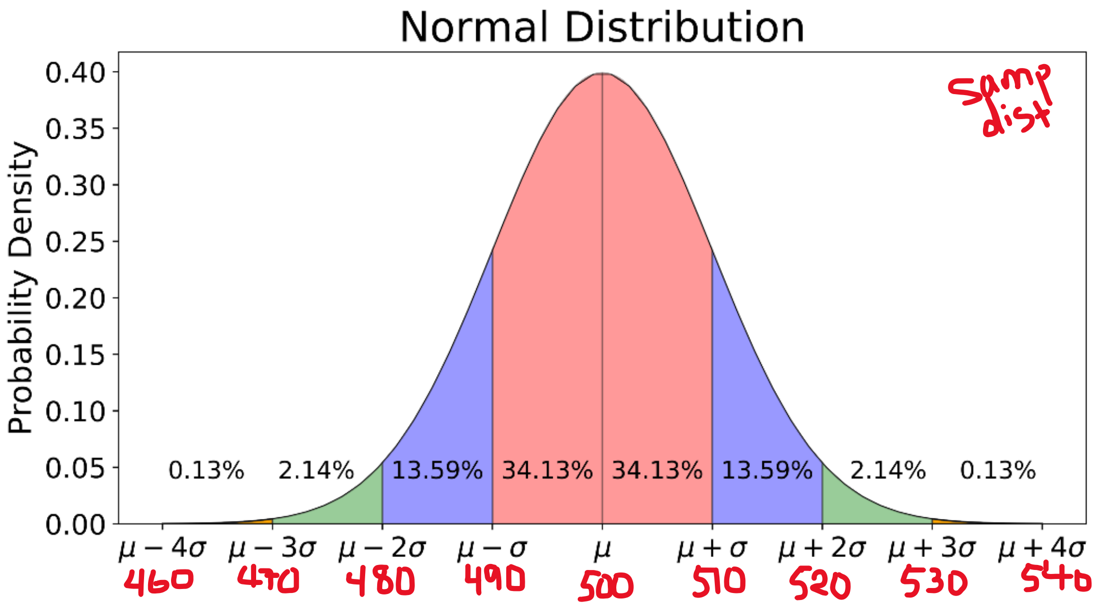
Suppose I take a sample of size n and measure the mean. $\bar{x}_1$
Then I repeat this: take another sample of the same size and measure the mean of that sample. $\bar{x}_2$
And over and over, say 1000 times.
μ_X is the mean of all these means.
I can plot those 1000 means in a distribution, in a histogram.
The standard deviation of those means in this distribution will be smaller than the population standard deviation, and it's smaller by a factor of the square root of the same size.
The distribution of the means is normal.

The idea here is that if we know how the mean is distributed (how it behaves when we take a lot of samples) then we can estimate how good the sample is without actually having to do all those repeated samples.

Sampling distribution: the distribution of the samples of means.
The standard deviation of the sampling distribution is called the standard error.

The first graph will be a graph of the population with mean μ and standard deviation σ. The second graph will be the sampling distribution for a sample of $n = 100$. The population has a mean of 500 and a standard deviation of 100.





$$\frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = \frac{100}{10}$$

$$= 10$$

In terms of our z-scores.

$$z = \frac{x - \mu}{\sigma}$$

This is the z-score for an individual observation
For a sampling distribution the x becomes $\bar{x}$, the mu becomes μ_X. And the standard deviation sigma is replaced with the standard error, $\frac{\sigma}{\sqrt{n}}$.
It becomes:

$$z = \frac{(\bar{x} - \mu_X)}{\frac{\sigma}{\sqrt{n}}}$$

Suppose I took a sample of 16 students, and I wanted to determine if the mean they obtained is unusual. The mean we got was 580. (This is an SAT test with mean of 500 and standard deviation of 100). What is their z-score? What is the probability of obtaining a mean of 580 or larger with this sample size?

Unusual: less than 5% likelihood (outside 2 standard deviations. Less than -2 or greater than 2, it is considered unusual.)

$$z = \frac{(\bar{x} - \mu_X)}{\frac{\sigma}{\sqrt{n}}} = \frac{(580 - 500)}{\frac{100}{\sqrt{16}}} = \frac{80}{\frac{100}{4}} = \frac{80}{25} = 3.2$$

So, this is an unusual average.

Probability calculation in Excel.

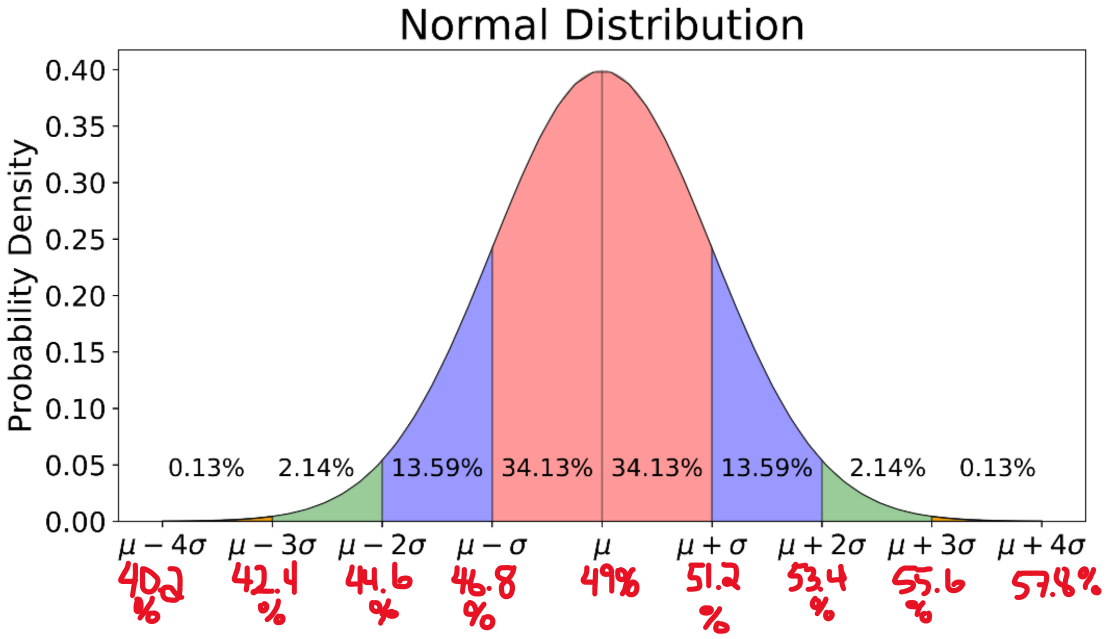
Calculate the probability exactly the same way that we did in Chapter 6. The only difference is the observation is a sample mean, and the standard error replaces the standard deviation.

Proportions

For proportions, the thought experiment is basically the same.
We take a sample of size n and calculate the proportion for a particular category. p_1
Then we take another sample of the same size, and find the proportion. p_2
And over and over again. Let's 1000 times.
Plot that in a histogram to look at the distribution of these proportions.
The distribution is also normally distributed. The mean of the distribution is the true proportion in the population, and the standard deviation is $\sqrt{\frac{p(1-p)}{n}}$ (the standard error formula).

We know that in the US population, men make up 49% of the population. That is our mean. If we take sample size of n=500. Draw the sampling distribution below.

Standard error $= \sqrt{\frac{0.49(1-0.49)}{500}} = 0.022356 \ldots \approx 2.2\%$



See Excel for an example. What is the probability of a company with 500 employees having fewer than 45% of the workforce being men?

Binomial distributions can often behave approximately like normal distributions (symmetrical, etc.)
Conditions that need to be met are
Low threshold: $npq \geq 5$
Better threshold: $npq \geq 10$

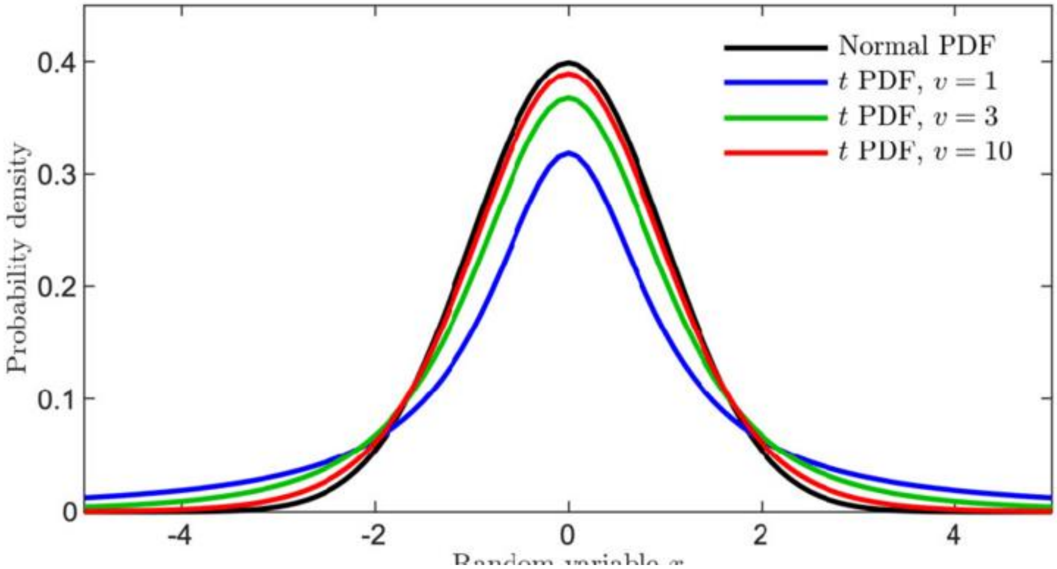
In our example: 500(0.49)(0.51)=124.95
If the proportion is smaller, say 10%, this number comes down.
500(0.1)(0.9)=45
What about 1%?
500(0.01)(0.99)=4.95 (this doesn't meet either condition)
If we increase the sample size to 1500
1500(0.01)(0.99) = 14.85 (this would meet the conditions)

The central limit theorem relies on the fact that the distribution of means (or proportions) gets more normal as the sample size increases. When the sample size is very large, the central limit theorem allows us to reliably use the normal distribution to estimate our probabilities.

For small sample sizes the difference between the distribution of means and the normal distribution is larger: the tails are more stretched out, and it only approximately is similar to the empirical rule. The center contains less of the samples, and the tails contain more.

In order to account for this fact, we use a slight different distribution to account for small sample sizes. In general, for sample sizes larger than 40 or so, it is relatively safe to use the normal distribution, especially if we have reason to think that the population is also normal.

For sample sizes less than 40 (or if the population is not normal, or just for a better estimates), we use the Student-T distribution.



We always use the normal distribution for proportions (assuming they meet the test above).

We will talk about specific situations when we talk about the confidence intervals for means in the next chapter.