The Central Limit Theorem

STA 198: Introduction to Health Data Science

Yue Jiang

June 05, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

A roadmap ahead

In just a few more lectures, we'll have learned enough tools to perform a pretty wide range of analyses and answer associated questions

In the latter half of this course, we'll emphasize specific scientific questions of interest, statistical methods for testing such questions, and translating those results to real-world conclusions

A roadmap for today

Today's topic may be a bit theoretical, but it is fundamental to making sound statistical inferences and conclusions

The basic idea is called the central limit theorem, which states that for **any** distribution with a well-defined mean and variance, the distribution of the means computed from samples of size n will be approximately normal (Gaussian)

What is statistical inference?

Statistical inference is the act of generalizing from a sample in order to make conclusions regarding a population while quantifying the degree of certainty we have

We are interested in population parameters, which we do not observe

Instead, we must calculate statistics from our sample in order to learn about the parameters

The sampling distribution of the mean

Suppose we're interested in the resting heart rate of students in STA 198, and are able to do the following:

1. Take a random sample of size n from this population
   - Calculate the mean resting heart rate *in this sample*, $\bar{x}_1$
2. Put the sample back, take a second random sample of size n
   - Calculate the mean resting heart rate from this new sample, $\bar{x}_2$
3. Put the sample back, take a third random sample of size n
   - Calculate the mean resting heart rate from this sample, too...

... ...and so on

After repeating this many times, we have a dataset that has the sample averages from the population: $\{\bar{x}_1, \bar{x}_2, \cdots\}$.

Can we say anything about the distribution of these sample means?

The central limit theorem

The central limit theorem states that for a population with mean μ and standard deviation σ, these three properties hold for the distribution of sample averages $\bar{X}$:

1. The mean of the sampling distribution is identical to the population mean μ
2. The standard deviation of the distribution of the sample averages is $\sigma/\sqrt{n}$, or the standard error (SE) of the mean
3. For n large enough (in the limit, as $n \to \infty$), the shape of the sampling distribution of means is approximately normal

What if the original population distribution is not normal?

The central limit theorem tells us that **sample averages** are normally distributed, if we have enough data. This is true **even if** our original variables are not normally distributed.

Interactive central limit theorem demonstration

Another experiment

Define a variable X to be 1 if a STA 198 student has brown eyes, and 0 if a student does not have brown eyes.

- What is the distribution of X?
- If we take a random sample, the average is an estimate of the true proportion of brown-eyed students in our population of interest
- If we take repeated random samples from our population and calculate the proportion in each sample with brown eyes, what values might we expect? Will we get the same values every time?

Another experiment

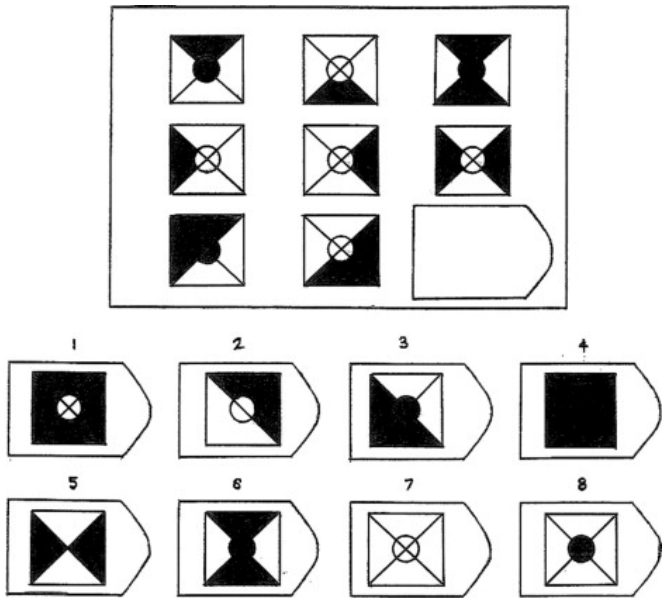
The central limit theorem tells us that the distribution of sample averages should have mean $E(X)$ and standard deviation $SD(X)/\sqrt{n}$

In our case, what are these values? Was this the case?

IQ tests

IQ tests are designed to have a probability distribution with $\mu = 100$ and $\sigma = 15$. Suppose we draw samples of size $n = 20$ from this population.

From the central limit theorem, the distribution of the sample averages will be approximately normal with mean 100 and standard deviation $15/\sqrt{20}$



IQ tests

If the population distribution is normal to begin with, then the distribution of the sample averages will also be exactly normal

If the population distribution is not normal, then the rule of thumb is that we need at least $n = 30$ for the central limit theorem to kick in for approximate normality

Example

Suppose I give a random sample of $n = 30$ STA 198 students an IQ test*, and the sample average score is 120. Does this mean that STA 198 students are smarter than average?

*I know there are lots of problems with IQ and IQ testing...bear with me here!

Example

The central limit theorem tells us that the distribution of means of samples of size 30 from this population is also normal, with mean $\mu = 100$ and $SE = \sigma/\sqrt{n} = 15/\sqrt{30} \approx 2.7$.

$Z = \frac{\bar{X}-\mu}{SE}$ is a standard normal random variable, and here $Z \approx 7.3$.

The probability of a z-score greater than this is extremely small. What does this mean?

Some additional questions

What are the upper and lower limits that enclose 95% of the means for samples of size n drawn from the population?

How large would our random samples need to be for 95% of their averages to lie within ± 10 of the population mean μ?