

Some prerequisite concepts

The exercises below cover a selection of topics you've likely seen in a previous statistics course(s). Try to solve them to the best of your ability.

1. In order to gauge public opinion about a referendum on the ballot in the upcoming elections in Indiana, a news organization conducted telephone interviews with 602 voters registered in the state. They found 74% of those interviewed supported the referendum.
 - a. Identify the population, the sample, the parameter(s), and the statistic(s).
 - b. Create a 90% confidence interval to estimate the true proportion of supporters.
2. A children's research organization studied the television viewing habits of 822 children between the ages of 8 and 14 years. They found the average time these children spent in front of a television per year was about 1054 hours. Identify the following: the population, the sample, the parameter, and the statistic.
3. Explain the meaning of a "sampling distribution." Use the previous two exercises as application context to illustrate the meaning of sampling distributions.
4. The distribution of annual income in a geographic region is skewed right, and has one mode. The overall mean and standard deviation (in thousands of dollars) are 45.8 and 36.2, respectively. What shape and summary statistics (i.e., mean and standard deviation) would we expect for the following distributions:
 - a. Incomes in a single random sample of size 1025 drawn from this population.
 - b. The sampling distribution of mean incomes in random samples of size 5.
 - c. Same as previous question, except random samples are of size 1025.

5. Suppose the normal distribution $N(700, 100)$ describes scores on a standardized test.
- What is the probability that a randomly selected individual from this population scores over 1000?
 - What is the probability the individual scores between 500-600?
 - There is only a 0.05 probability that the individual will score more than how much?
 - What is the probability the individual scores exactly 700?

The Central Limit Theorem

Why you should care:

1. It is the absolute, total, complete, comprehensive, theoretical foundation of almost all research in the social sciences, business & economics, certain areas of the natural sciences, and many other data-centric fields.
2. Unlike most other areas of math, this single theorem alone is sufficient for understanding most strategies used in statistical research (including in Data Science).

What the CLT says:

When the conditions are met, the sampling distribution of mean values follows the normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$, where n =sample size, μ =population mean, and σ =population standard deviation.

The conditions that must be met:

1. The sample size must be large enough.
2. The sample must be made up of independent observations (or cases).

Remark: There are many nuggets of wisdom contained here beyond just what you see on the surface – e.g., does shape of population distribution matter? what is the effect of sample size? what sampling methods are appropriate for use in our study?