The Sample Mean / IID setting

Reading

- Section

Practice Problems

. (Page) 41, 42
. (Page) 86, 89, 95

Notes

The IID setting

The IID setting is somewhat analogous to the binomial setting in the case where the values in question are scalar. Here are its characteristics:

1. Fixed number of trials. As in the binomial setting, we are repeating something a fixed number of times. We will again denote that by n. It is often called the *sample size*, as it is usually exactly what it is.
2. Trials have numerical values as outcomes. As such we can describe them as random variables. for instance we might be selecting students at random and looking at their gpa. We will denote by X_1, X_2, X_3, ..., X_n the random variables for each of the trials.
3. The distributions of the different trials are all identical. In other words, the "tables" for X_1, X_2, and so on are all identical. We often use X to denote that common distribution. In the example with the GPAs, this means that the kinds of values we can get when we look at the GPA of the first student we pick are the same as the kinds of values we can get when we look at the second student we pick and so on. In a sample case, this basically means that *all samples are drawn from the same population*.
4. The trials are independent of each other. In our example, it would mean that what gpa we get for the first say 5 students does not have an effect on the gpa we might get for the 6th student. The checks we need to perform for this independence are the same checks we had to perform to determine if the trials in a binomial are independent. In particular, in the case where we have an actual population and removing people from it to form the sample, then we need to know that the population is at least 20 times the sample size.

We can summarize this by saying:

> In the IID setting we have a *fixed number of **I**ndependent, **I**dentically **D**istributed trials*

You should see a lot of similarities with the binomial. In the case of the binomial we had the same chance of success, p. The analog of this here is the claim that the distributions of all trials are identical.

Sample Mean

In the IID setting, the quantity of interest is the **sample mean**:

$$\bar{x} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Notice that it is a random variable, the sum of all the X_i. Its value depends on the sample we end up with, just like the value of $\hat{p}$ depended on the sample in the binomial case. So *different samples would give us different values for the sample mean.*

Just like in the binomial we were interested in the kinds of values that $\hat{p}$ can take, and how likely each is, we can do the same thing here.

Sampling Distribution

The **sampling distribution** of $\bar{x}$ is the distribution of the values that $\bar{x}$ takes across all possible samples of size n.

The remarkable fact is that we can describe what this distribution is, even if we know very little about the values that the X_i can take. Let us set up the stage.

In the IID setting we draw n samples/trials from a distribution X. We will denote the mean of that distribution by μ and the standard deviation of this distribution by σ.

Then we can compute the mean and standard deviation of the *sampling distribution of* $\bar{x}$:

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The greek letters on the left-hand side denote the mean and standard deviation of the random variable $\bar{x}$, in other words the mean and standard deviation of the sampling distribution.

Let us rephrase this:

Suppose we draw samples of size n by drawing independent values from a population X with mean μ and standard deviation σ.

If we then compute the sample mean values $\bar{x}$, one for each possible sample of n values, then the mean of these values is μ and the standard deviation is $\frac{\sigma}{\sqrt{n}}$.

Sample averages vary less than the original values, by a factor of $\sqrt{n}$.

Sample averages are on average the same as the original values.

This tells us at least the mean and standard deviation of the sampling distribution of $\bar{x}$. Amazingly, we can say more about it. This is the famous Central Limit Theorem.

The Central Limit Theorem

Central Limit Theorem

When the sample size n is "sufficiently large", then the sampling distribution of $\bar{x}$ will be approximately normal.

So we can assume that $\bar{x}$ follows the distribution:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

This is a remarkable theorem. It tells us that no matter what kind of distribution our original values had, heavily skewed, outliers, multiple modes and so on, then once we take large enough samples, the possible values are behaving like a normal distribution. *No matter what we started with.*

This is the reason why we have standardized tests. Your score in a standardized test is an average of your scores in many questions, and averages tend to behave in a more normal way than the original values.

- *Sample averages are on average the same as the original values.*
- *Sample averages vary less than the original values.*
- *Sample averages are more normal than the original values.*

The only thing left is to answer the question of what is "sufficiently large sample size". All we have is a general rule of thumb, but the bottom line is: *The more non-normal the original population, the larger the sample size you would need.*

Rule of Thumb for sufficient sample size for the Central Limit Theorem to apply.

- If the original population is heavily skewed, outliers etc, we would need a sample size near $n = 100$ or more.

- If the original population is only slightly skewed, without many outliers, a sample size around $n = 40$ or more would suffice.
- If the original population is close to symmetric, a sample size of 10-20 is enough.
- If the original population is normal, then even a sample size of $n = 1$ is sufficient.

The larger the sample size, the better. These are starting points depending on the population.

One important observation is that these sample sizes are just the minimums required to be able to claim that $\bar{x}$ is normally distributed. We typically still need even bigger sample sizes, in order to keep $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ small.

An example

Suppose we draw at random a sample of size 40 from the Hanover student body, and consider their GPAs.

- We first consider whether this fits into the IID setting.
  - The students are selected at random, so that's a good start.
  - We do remove the students from the population as we select them, so we need to make sure we have at least 20 times as many students to begin with. Since $20 \times 40 = 800$, as long as we have more than 800 students on campus we are OK with the assumption of independent trials.
  - Each student we pick at random has the same possible outcomes for their GPAs. This would not be the case for instance, if we constantly alternated between selecting a male student and a female student, as in general those two groups have different GPA distributions. But if we trully pick the students at random, then they would be identically distributed.
  - We do not know what the mean and standard deviation of the entire population is, but that's what μ and σ would stand for. So μ is the average GPA of students at Hanover, and σ is the standard deviation of all student GPAs.
    - For the purposes of this problem we will assume that the average Hanover GPA is $\mu = 2.98$, and that the standard deviation of Hanover GPAs is $\sigma = 0.55$. The true value for σ might be a bit smaller, but it is unlikely to be much larger (the allowed range of GPAs prohibits it).
  - All these allow us to say we are in an IID setting, with $n = 40$. And $\bar{x}$, is the average GPA in the sample of 40 we selected.
- Next we need to consider if we can apply the Central Limit Theorem.
  - In order to answer that, we would need to know something about how the population is distributed.

- Our sample of $n = 40$ fits into the "second" option in the list earlier. In other words, if the population distribution is NOT worse than slightly normal, we are OK.
- We don't actually know the population distribution. We have to therefore make an educated guess. In this instance we have two tools at our disposal.
  * Prior knowledge. College GPAs in general tend to approach a normal distribution.
  * Examine the sample: We have 40 values in our sample, and they came from the population. If the population was heavily skewed, then it is reasonable to expect that we would be able to observe some of that skeweness in our sample values. Therefore plotting a histogram of the 40 values in our data would be a good place to start. If we don't see much skeweness there, we can guess that the population didn't have much skewness either.
- So this would allow us to say that the Central Limit Theorem "kicks in". Therefore $\bar{x}$ is distributed normally.

- Next, we compute the mean and standard deviation of this distribution.
  - $\mu_{\bar{x}} = \mu = 2.98$
  - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.55}{\sqrt{40}} = 0.087$
  - Notice how considerably smaller this standard deviation is: Our sample mean $\bar{x}$ cannot deviate all that much from the 2.98 value.

- All this leads us to being able to say that $\bar{x}$ behaves according to the normal distribution $N(2.98, 0.087)$. We can use our knowledge of that distribution to answer various questions about $\bar{x}$.

  For example, say I want to find a range that captures 90% of the possible $\bar{x}$ values.
  - In terms of the normal distribution, I need to leave a 5% out on either side. So we will be working with $p = 0.05$.
  - Table A tells us that the corresponding z value is $z = -1.645$.
  - We scale that back to an $\bar{x}$ value using the formulas

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$
$$\bar{x} = \sigma_{\bar{x}} \cdot z + \mu_{\bar{x}}$$

  - In our case that means $\bar{x} = \pm 0.089 \cdot (-1.645) + 2.98 = 2.98 \pm 0.1464$.
  - The two values we get that way are 2.83 and 3.13.
  - So what this means is that 90% of the possible samples out there have a sample mean value $\bar{x}$ somewhere between 2.83 and 3.13. So 90% of the times when we choose a sample we're no more than 0.15 away from the actual mean.