Sampling Distributions

Grinnell College

January 24, 2024

Review

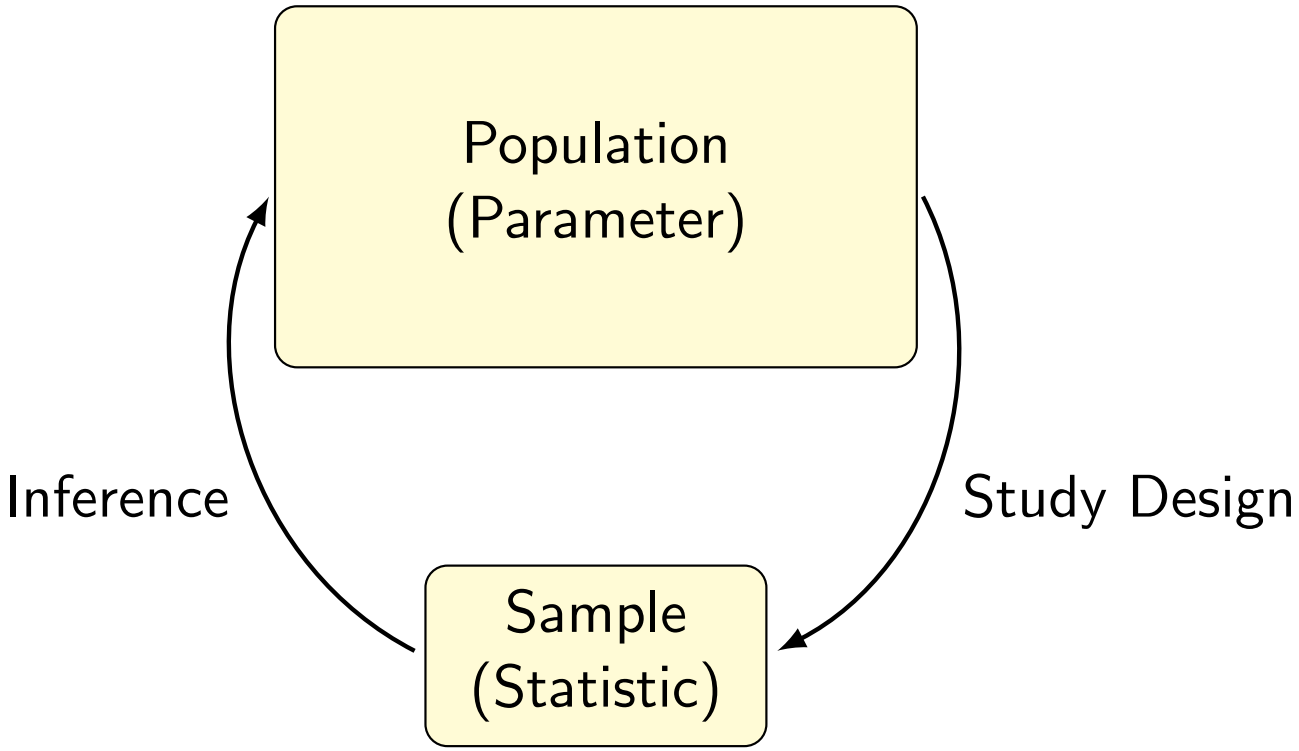
Last week, we considered randomness and distributions

- A distribution describes relationship between "events" and probabilities
- Sampling is a *random process*
- Central Limit Theorem (CLT) tells us that sample mean follows a normal distribution

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Specifically, this tells us that the sampling distribution of $\overline{X}$ has an expected value $E(\overline{X}) = \mu$ and standard deviation $\hat{\sigma} = \sigma/\sqrt{n}$

The Statistical Framework



Samples are Random ($n = 20$)

Each random sample will have a different (random) sample mean, $\overline{x}$

Review of Sampling Distribution

Recall last time that, as a consequence of the CLT, we made the claim that by only collecting a single sample from a population, we would be able to approximate its distribution and leverage this to make further claims regarding our statistic

Specifically, we noted that our statistic follows a *normal distribution*, centered around the true population mean, with an estimate of variability based on sample size

Our intention here is to construct a suitable interval of values around the sample mean that contains the true mean with some specified probability

Normal Distribution

A quick reminder on some properties of the normal distribution

Simulation and Statistics

In some sense, accepting that a single sample yielding a single estimate of the mean and standard deviation can follow a particular distribution demands a leap of faith. Let's try and make this leap shorter

To do so, we will approach this same problem from two different ways: through the use of simulation and an application of the CLT (admittedly, also via simulation). Arriving at the same conclusions in each should give us confidence that the methods are equivalent

This is especially handy, considering that only one of them can be used for practical purposes

Notation

We will have some unfortunately overlapping notation, so this slide will serve to be a reference when reviewing

1. $\mu =$ population mean and $\sigma =$ population s.d.
2. $\overline{x}_i$ will be sample mean from ith sample
3. $\hat{\sigma}_{\overline{x}}$ will be the standard devition of 1,000 samples of $\overline{x}_i$
4. $\hat{\sigma}_n$ will be the standard deviation from a single sample of size n

Simulation

For our simluation, we will follow these steps

1. Start with a population with mean $\mu = 0$ and standard deviation $\sigma^2 = 1$
2. We will collect 1,000 samples of size $n = 25$ and for each one, compute the sample mean $\overline{x}_i$
3. We will plot all 1,000 samples of $\overline{x}_i$, creating a histogram, allowing us to visualize the resulting distribution
4. We will confirm that the expected value is $E(\overline{X}) = \mu = 1$ and that the standard deviation is $\hat{\sigma}_{\overline{x}} = \sigma/\sqrt{n} = 0.2$
5. Finally, we will look at the interval $\overline{x} \pm \hat{\sigma}_{\overline{x}}$ and confirm that it contains about 68% of the total observations

Simulation

Under CLT, the distribution of $\overline{X}$ should have mean 0, standard deviation of 0.2, with 68.2% of observations between $\overline{X} \pm \hat{\sigma}$

- Average value of sample is $\overline{x} = -0.00622$
- Standard deviation of sample is $\hat{\sigma}_{\overline{x}} = 0.19104$
- Interval $\overline{x} \pm \hat{\sigma}_{\overline{x}}$ contains 68.9% of total observations

Statistics

For our verification with statistics, we will do something slightly different:

1. Start with the same population with $\mu = 0$ and $\sigma^2 = 1$
2. We will collect 10 samples of size $n = 25$, and for each one, we will compute the sample mean $\overline{X}$ and standard error $\hat{\sigma}$
3. For each sample, we will look at the interval $\overline{x} \pm \hat{\sigma}$ and compare it with what we saw in the simulation

Statistics

Here, 60% of the constructed confidence intervals contain the true population mean, $\mu = 0$

Sample	$\overline{x}_n$	$\hat{\sigma}_n$	$\overline{x}_n \pm \hat{\sigma}_n/\sqrt{n}$
1	0.28	0.96	(0.09, 0.47)
2	0.52	1.04	(0.31, 0.72)
3	0.03	1.32	(-0.24, 0.29)
4	-0.55	1.05	(-0.76, -0.34)
5	0.13	0.75	(-0.02, 0.28)
6	-0.06	0.94	(-0.25, 0.12)
7	0.01	1.1	(-0.21, 0.23)
8	-0.13	0.98	(-0.32, 0.07)
9	-0.09	1.36	(-0.36, 0.19)
10	-0.22	0.96	(-0.41, -0.03)
Average	-0.008	1.04	(-0.217, 0.201)

What does this mean?

The full simulation portion above allowed us to actually perform iterations of this random process – by carrying this process out and examining the results, we were able to confirm empirically that the observed distribution was as expected

The "statistics" portion enabled us to really zoom in on ten of the samples collected to see what would happen if, using that sample alone, we made an estimate of mean, along with an interval about that mean

While the individual estimates and intervals themselves showed some variability, *on average*, they agreed with what was found in the full simulation

Intervals

For each of the methods investigated, we concerned ourselves with the construction of the interval $\overline{x} \pm \hat{\sigma}$ which, according to properties of a normal distribution, should contain roughly 68.2% of the total observations

This bore out in the simulation, with the constructed interval containing 68.9% of the total simulated sample means

And in the second portion, we nearly found this, with 60% of the constructed intervals containing the true mean $\mu = 0$. Had we examined more than ten samples, this proportion would have become increasingly closer to 68.2%

As it turns out, the process behind both of these constructions is identical

Intervals, cont.

Let's limit our attention for now on the ten intervals we constructed, where 60% of them did not contain the true population mean. This value is known as the *coverage probability*

Critically, the coverage probability is associated with *the random process of generating intervals*, **not** the probability that a particular interval contains the true parameter value

In other words, a particular interval either does or does not contain the true parameter value. We don't know, and we have no way of knowing for sure. We can only make probablistic statements about the process itself

Review

1. Sampling distribution as random process
2. Coverage probability describes probability that a *process* of constructing intervals contains true parameter
3. Wider intervals have higher coverage probability but are also less informative