Wednesday, Feb 23

The Number of Observational Units

Let N denote the number of "observational units" in a population of observations. This can be infinite or finite.

- In an *experiment* the number of observational units is (theoretically) *infinite*.
- In an *survey* the number of observational units is *finite*.

When the number of observational units is *finite* we need to be concerned with if we are sampling *with* or *without* replacement.

Sampling With Replacement

When **sampling *with* replacement**, it is possible to observe the same unit more than once.

Example: Suppose we had a very small forest with $N = 3$ trees with volumes of 10, 20, and 30 cubic feet. Assume we will obtain a sample of $n = 2$ observations of tree volumes using sampling *with* replacement. We select trees one at a time, and each time we do every tree in the population has the same probability of being selected.

Table 1: Sample Space

Sample	Probability	$\bar{x}$
10,10	1/9	10
20,10	1/9	15
30,10	1/9	20
10,20	1/9	15
20,20	1/9	20
30,20	1/9	25
10,30	1/9	20
20,30	1/9	25
30,30	1/9	30

Table 2: Sampling Distribution

$\bar{x}$	Probability
10	1/9
15	2/9
20	3/9
25	2/9
30	1/9

Here $\bar{x}$ has mean 20 and standard deviation (i.e., standard error) of about 5.77.

Sampling Without Replacement

When **sampling *without* replacement**, it is *not* possible to observe the same unit more than once.

Example: Suppose we had a very small forest with $N = 3$ trees with volumes of 10, 20, and 30 cubic feet. Assume we will obtain a sample of $n = 2$ observations of tree volumes using sampling *without* replacement. We select trees one at a time, and each time we do every tree in the population *that was not previously selected* has the same probability of being selected.

Table 3: Sample Space

Sample	Probability	$\bar{x}$
10,20	1/6	15
20,10	1/6	15
10,30	1/6	20
30,10	1/6	20
20,30	1/6	25
30,20	1/6	25

Table 4: Sampling Distribution

$\bar{x}$	Probability
15	1/3
20	1/3
25	1/3

The sampling distribution shows that $\bar{x}$ has mean 20 and standard deviation (i.e., standard error) of about 4.08.

Example: Now consider another forest with $N = 100$ trees, of which 60 have a volume of 10 cubic feet, and 40 have a volume of 20 cubic feet, and suppose we select a random sample of $n = 5$ trees by using sampling *without* replacement.

Table 5: Population Distribution

Volume	Frequency	Probability
10	60	0.6
20	40	0.4

	Probability		
Observation	10	20	Sample
1	60/100	40/100	10
2	59/99	40/99	10, 20
3	59/98	39/98	10, 20, 10
4	58/97	39/97	10, 20, 10, 10
5	58/96	38/96	10, 20, 10, 10, 20

If we were sampling *with* replacement, the probability of the sample 10, 20, 10, 10, 20 would be

$$0.6 \times 0.4 \times 0.6 \times 0.6 \times 0.4 = 0.03456,$$

but because we were sampling *without* replacement, the probability of the sample is

$$0.6 \times 40/99 \times 59/98 \times 58/97 \times 39/96 \approx 0.0347294.$$

When sampling *without* replacement, the observations are *not independent*. Observations are independent only if the probabilities for one observation *do not depend on the other observations*.

Inferences for μ When Sampling Without Replacement

Suppose we are estimating μ with $\bar{x}$. The standard error formula $s/\sqrt{n}$ assumes sampling *with* replacement. When sampling *without* replacement the correct formula for the standard error of $\bar{x}$ is

$$\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}},$$

where N is the *population size*. The margin of error is then

$$t\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}},$$

and the confidence interval for estimating μ is

$$\bar{x} \pm t\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}.$$

The term $1 - n/N$ is sometimes called the **finite population correction** (FPC).

Example: The mean test score of a simple random sample *without replacement* of 50 students from a school of 200 is 70 points, with a standard deviation of 10 points. What is the margin of error and confidence interval for the mean test score for all 200 students?

Example: Conservation researchers conducted a survey to estimate the density and abundance of a species of "larkspur" (a flowering plant of the genus *Delphinium*) in a rectangular region. To survey the larkspur they divided the region into 150 one-meter wide transects from north to south. The researchers selected a sample of 20 transects using random sampling. The mean density of larkspur in these 20 transects was 0.83 larkspur per square meter, and the standard deviation was 0.72 larkspur per square meter. What is the point estimate, margin of error, and confidence interval for the *mean density of larkspur in the region*?

Question: In some cases we might want to omit the finite population correction term $1 - n/N$ by effectively setting $1 - n/N = 1$ so that $\sqrt{1 - n/N}$ "disappears" from our equations. Sometimes we do this to simplify calculations, but also we may need to do this if we do not know N. When would it be reasonable to omit the finite population correction term?

Estimating a Population Total (τ)

In survey sampling, the parameter μ is the *mean* of *all* units in the population, so that

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

The parameter τ is the *total* of *all* units in the population, so that

$$\tau = x_1 + x_2 + \cdots + x_N.$$

Note that $\mu = \tau/N$ and $\tau = N\mu$.

Examples of population totals that we might want to estimate:

1. The total volume of N trees in a forest.
2. The number of trees in N transects through a forest.
3. The number of pottery fragments in N grid squares over an archaeological site.
4. The number of mental health counseling/therapy visits made by N students or employees.
5. The total expenditure on phone apps by N consumers.

The point estimate of τ is $N\bar{x}$ since $\tau = N\mu$ and $\bar{x}$ estimates μ. The standard error of $N\bar{x}$ (assuming sampling without replacement) is

$$N\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}},$$

so the margin of error for estimating τ is

$$tN\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}},$$

and the confidence interval for τ is

$$N\bar{x} \pm tN\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}.$$

Note: We will always assume sampling without replacement when estimating τ.

Example: Based on a random sample (without replacement) of $n = 100$ students at a university of $N = 1000$ students, the mean expenditure per year on phone apps was $\bar{x} = 20$ dollars with a standard deviation of $s = 10$ dollars. What is the estimate of the *total* expenditure on phone apps per year for all of the students at the university?

Example: Conservation researchers conducted a survey to estimate the density and abundance of a species of "larkspur" (a flowering plant of the genus *Delphinium*) in a rectangular region. To survey the larkspur they divided the region into 150 one-meter wide transects from north to south. The researchers selected a sample of 20 transects using random sampling. The mean number of larkspur in these 20 transects was 21.45 larkspur, and the standard deviation was 21.52 larkspur. What is the point estimate, margin of error, and confidence interval for the *total number of larkspur in the region*?