Chapter 8: Sampling Variability and Sampling Distributions


Sampling distributions

Three distributions: population, data, sampling

Sampling distribution of the sample proportion

Sampling distribution of the sample mean

---

**LLN and CLT**

*LLN*: $\overline{X}_n \rightarrow \mu$ as $n$ gets larger.

*CLT*: the distribution of $\overline{X}_n \rightarrow \text{Normal Distribution}$ as $n$ gets larger.

Both theorems require certain conditions to be satisfied for the theorem to be applicable. The CLT requires

- independent sample elements (from less than 10% of the population)
- relatively large sample size (for instance, at least 30 elements)
- data which are not highly skewed and without extreme outliers

in order to conclude that the sample mean is approximately normally distributed with standard deviation approximately SE (OIS, p.168).
Sampling distributions of the sample mean

Uniform distribution.

Exponential distribution.
Normal distribution. The sampling distribution of a normal distribution is a normal distribution, even for small sample sizes.

Sampling distributions of counts and proportions

Imagine taking a sample of size 100 from a Bernoulli population with $p = 0.6$. Do this 1000 times and make a histogram of the counts of successes. Take another 1000 samples of size 100 and make a histogram of the proportions of successes in the samples. What do you observe?

Distribution of a statistic (mean or proportion)

Two scenarios: (1) If we are studying a quantitative variable and we know the population parameters $\mu$ and $\sigma$, what can be said of the sample statistics? (2) If we know the sample statistics $\bar{x}$ and $s$, what can be said of the population parameters? The first question is easiest to answer, but is rarely the case. The second question leads to much of contemporary statistics.
We might be studying a quantitative variable in a population (normal or not) with population parameters \( \mu \) and \( \sigma \), and we wish to know the sampling distribution of the sample mean \( \bar{x} \). Or we might be studying a categorical variable in a population with proportion \( p \), and we wish to know the sampling distribution of the sample proportion \( \hat{p} \). The following table summarizes the sample distributions in both cases (Probability and Statistics, Open Learning Initiative, CMU).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistic</th>
<th>Shape</th>
<th>Center</th>
<th>Standard Error</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>quantitative (( \sigma ) known)</td>
<td>( \bar{x} )</td>
<td>Normal</td>
<td>( \mu )</td>
<td>( \sigma/\sqrt{n} )</td>
<td>( n \geq 30 ) or approx. normal</td>
</tr>
<tr>
<td>quantitative (( \sigma ) unknown)</td>
<td>( \bar{x} )</td>
<td>( t )</td>
<td>( \mu )</td>
<td>( s/\sqrt{n} )</td>
<td>( n \geq 30 ) or approx. normal</td>
</tr>
<tr>
<td>categorical</td>
<td>( \hat{p} )</td>
<td>Normal</td>
<td>( p )</td>
<td>( \sqrt{\frac{p(1-p)}{n}} )</td>
<td>( \min(np, n(1-p)) \geq 15 )</td>
</tr>
</tbody>
</table>

**Barrels of marbles and beans**

*Three distributions: population, data, sampling*

Here is a more leisurely view of sampling distributions. Suppose that in the middle of the classroom we have a big barrel of red and white marbles. A proportion \( p \) of this large population of marbles is red. If we actually knew exactly how many red and how many white marbles were in the barrel, we could make a small table displaying the number of red and white marbles in the barrel. Agresti and Franklin would call this table the *population distribution*. Beside that barrel is another large barrel full of bright green lima beans. The beans vary in size from large to small, but most are medium-sized. We are actually interested in the weights of the beans, so it is interesting to know that the average weight of the beans in this barrel is \( \mu \) and the standard deviation of their weights is \( \sigma \). A graph of the actual distribution of weights of the lima beans in the barrel would be an illustration of a population distribution.

**Sampling distribution of the sample proportion**

Next to the barrels is a small scoop. We begin by sampling from the barrel with the marbles. Scoop out a small but fixed number of marbles, \( n \), and count the proportion of red marbles in your scoop, \( \hat{p} \). Make a small table displaying the number of red and the number of white marbles in your scoop. Agresti and Franklin would call this table the *data distribution* of the marbles in your sample. Another student takes a scoop of the same number \( n \) of marbles, and calculates the statistic \( \hat{p} \) for the marbles in her scoop. We might expect these two statistics to be similar, but probably not exactly equal. The question now is, if a large number of students took similar samples of size \( n \) from the barrel of marbles, and calculated the proportion \( \hat{p} \) of red marbles in each scoop, what would the collection of statistics \( \hat{p} \) look like. The answer to that question is called the *sampling distribution* of \( \hat{p} \). Since our samples are independent, the CLT says that the sampling distribution is approximately normal. It has mean \( \mu \) and standard deviation \( \sqrt{\frac{p(1-p)}{n}} \).

**Sampling distribution of the sample mean**

Now turn to the barrel of lima beans. We want to study the weights of the beans in our samples, so we borrow a precision balance from the chemistry department. Scoop out a small but fixed number of lima beans, \( n \), weigh each bean in your scoop, and calculate the average weight of the beans in your scoop, \( \bar{x} \). Make a stripchart (dot plot) of the weights of the beans in your sample. This strip chart is an illustration of the *data distribution* of the beans in your sample. Now the next student repeats the process, and calculates the average weight of the beans in the second scoop, \( \bar{x} \). We might expect these two statistics to be similar, but probably not exactly equal. The question now is, if a large number of students took similar samples of
size $n$ from the barrel of lima beans, and calculated the average weight of the beans $\bar{x}$ in each scoop, what would the collection of statistics $\bar{x}$ look like. The answer to that question is called the sampling distribution of $\bar{x}$. Since our samples are independent, the CLT says that the sampling distribution is approximately normal, and in fact if the distribution of the population weights is actually normal, then the sampling distribution is also normal (not just approximately normal). In either case, the sampling distribution has mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.

**Standard Error**

The standard deviation of a statistic is called a standard error, so in the sequel we will write

$$SE_{\bar{x}} = \sigma/\sqrt{n}$$

or

$$SE_{\hat{p}} = \sqrt{p(1-p)/n}.$$  

---

**Cherry Blossom Run, 2012**

16,924 runners participated in the Cherry Blossom 10 Mile Run in 2012. Take the population to be all of the runners, and estimate their average time to complete the race and the average age of the participants by taking 1000 samples of 100 runners each and calculating the corresponding sample statistics. The results are illustrated in the population and sampling distributions shown below. (OpenIntro Statistics, 2nd ed., pp.159-164).
Inference

Outline for one-variable inference for sample data (Peck, chapters 7–13):

The goal is to generalize from a sample to learn about a population

- categorical variable
  - one proportion
    - confidence interval - one-sample $z$ CI for a proportion
    - hypothesis testing - one-sample $z$ HT for a proportion
  - difference of two proportions
    - confidence interval - two-sample $z$ CI for a difference in proportions
    - hypothesis testing - two-sample $z$ HT for a difference in proportions

- continuous variable
  - one mean
    - confidence interval - one-sample $t$ CI for a mean
    - hypothesis testing - one-sample $t$ HT for a mean
  - difference of paired means
    - confidence interval - paired $t$ CI for a difference in means
    - hypothesis testing - paired $t$ HT for a difference in means
  - difference of independent means
    - confidence interval - two-sample $t$ CI for a difference in means
    - hypothesis testing - two-sample $t$ HT for a difference in means

Exercises

We will attempt to solve some of the following exercises as a community project in class today. Finish these solutions as homework exercises, write them up carefully and clearly, and hand them in at the beginning of class next Friday.

Homework 8a – sampling distributions

Exercises from Chapter 8:
8.2 (histograms), 8.3 (imports), 8.15 (sampling distribution), 8.16 (normal), 8.26 (sampling distribution)

Homework 8b – sampling distributions

Exercises from Chapter 8:
8.27 (sampling distribution), 8.29 (polling), 8.30 (credit card), 8.50 (sample size), 8.52 (hurricane)