Inference in the Sampling Model

Sep 28 2015
Announcements
Today

A recap of statistical inference in the random sampling model

Wednesday: paired t-test
Your turn

What terms did you highlight?
An example of the statistical process in the random sampling model

We have a question.

Do textbooks cost more at the bookstore than on Amazon?

We translate this to a question about a population distribution. What is the mean difference in price between Amazon and the bookstore, for all textbooks required in OSU classes? Is this mean bigger than zero?

We can’t/won’t/don’t collect data on the whole population, but instead get a sample from the population and use properties of the sample to estimate properties of the population.

The average price difference in our sample of size, n=100, is $10 with a sample standard deviation of $5. With 95% confidence we estimate that OSU textbooks on Amazon cost between $9 and $11 less than at the bookstore.
Random sampling model

Single population

Population

Every member has a number associated with it

Bookstore price - Amazon price = $62 - $57 = $5

population differences

distribution of Bookstore price minus Amazon price for all OSU books

Sample

distribution of Amazon price minus bookstore price for our sample of OSU books

chem 101
Amazon price: $89
Bookstore price: $91
Difference: $3

jane eyre
Amazon price: $7
Bookstore price: $13
Difference: $5

intro bio
Amazon price: $124
Bookstore price: $123
Difference: -$1

where's the center?
Histograms and distribution functions

A histogram is a graphical representation of the distribution of a finite set of numbers. To find the number of observations that were in a given range we add the heights of the bars.

Sometimes you’ll see a smooth curve as a representation of a population distribution. Think of it like a histogram with infinitely small bin width, and infinitely many observations. Areas under the curve represent probabilities.

Technically, this is called a probability density function, ST521.
Statistical Inference

*Population inference* is using a sample to infer properties of the population.

*For the textbooks:* using the sample average to infer the population mean

This is **statistically justified** as long as:

- observations are **sampled at random** from the **population of interest**.

Chance enters the study through the act of randomly taking a sample. This is one of two “mechanisms of chance” we will cover.
The key to making inferences in the random sampling model is the relationship between the population distribution and the sampling distribution.

What is a sampling distribution?
# Parameters, Statistics and Estimates

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
<td>An <strong>unknown</strong> value in a probability model</td>
<td>Population mean, $\mu$ (mu)</td>
</tr>
<tr>
<td>(of a population,</td>
<td></td>
<td>Population standard deviation, $\sigma$ (sigma)</td>
</tr>
<tr>
<td>or of a model)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Statistic</strong></td>
<td>Something you can calculate from data</td>
<td>Sample average, $\overline{Y}$ (y bar) or $\overline{X}$</td>
</tr>
<tr>
<td>(of a dataset)</td>
<td></td>
<td>Sample standard deviation, SD or s</td>
</tr>
<tr>
<td><strong>Estimate</strong></td>
<td>A statistic used as a guess for a parameter</td>
<td>The sample average is an estimate of the population mean.</td>
</tr>
<tr>
<td>(of a parameter)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sampling Distribution

**Sleuth:** histogram of all values for the statistic from all possible samples that can be drawn from a population

**Nature Article:** Sample parameters have their own distribution called the sampling distribution, which is constructed by considering all possible samples of a given size.

**OpenIntro:** distribution of the point estimates based on samples of a fixed size from a certain population.
**a** Here’s a population

**b** Imagine we take a sample of size n=5 from this population. One example would be {1, 9, 17, 20, 26}, it’s sample average is 14.6. But that is only one possible sample.

**c** Imagine all the other possible samples. For each sample find it’s sample average and make a histogram of these sample averages. This is the sampling distribution of the sample average.
Figure 3

**Population distribution**

- **Normal**
- **Skewed**
- **Uniform**
- **Irregular**

**Sampling distribution of sample mean**

- *n* = 3
- *n* = 5
- *n* = 10
- *n* = 20

The CLT tells us that the distribution of sample means (indicated by vertical dotted lines) will become increasingly close to a normal distribution as the sample size increases, regardless of the shape of the population distribution. This relationship is captured by one of the most important and omnipresent concepts in statistics: the sampling distribution.

The omnipresence of variability will ensure that each sample will be different from the population, even if the sample size is quite large. However, it turns out to be an extremely useful concept in our estimate of the population mean with increase in sample size. Notice that it is still possible for a sample mean to fall far from the population mean, especially for small samples. However, as the sample size increases, the sampling distribution of the sample mean gets closer and closer to the population distribution. The terms in the second relationship are the spread of sample means, and its spread is the s.e.m. The sample mean and s.d. (indicated as in sample means for 10,000 samples) are used to estimate those of the population. The sample mean and s.d. are denoted by *X̄* and *s*. The distinction between sample and population parameters is emphasized by the use of Roman letters for samples and Greek letters for populations.

As we increase the number of samples, the number of times the sample mean falls outside the interval of two standard errors of the mean (2 s.e.m.) from the population mean, especially for small samples, will not change (sample size increases, regardless of the shape of the population distribution). As we increase the number of samples, the number of times the sample mean falls outside the interval of two standard errors of the mean will decrease (our estimate of the population mean with increase in sample size draws from the distribution in the CLT, the precision increase of a sample's estimate of the population mean, especially for small samples). Just like the population, the sampling distribution is not directly measurable because we do not have access to all possible samples. However, it is free from bias. For example, surveys sample only individuals who agreed to participate and do not capture information about those who refused. These two groups may be meaningfully different.

The CLT tells us that the distribution of sample means (for example, Fig. 2a) is quite a bit smaller than that of the population, *s.d. (skew)*, which is very different than the population in *X̄* and *s*.

Note: Any Supplementary Information and Source Data files are available in the online version of this article.
**Facts** about the sampling distribution for the sample average

Regardless of the shape of the population distribution, the sampling distribution:

1. will have the same mean as the population distribution \( \mu_X = \mu \)
2. have a smaller standard deviation \( \sigma_X = \frac{\sigma}{\sqrt{n}} \)
3. and it’s shape will be closer to a Normal distribution than the population distribution (how close depends on the sample size and how close the population distribution was to Normal).

**Central Limit Theorem**
The key to making inferences in the random sampling model is the relationship between the population distribution and the sampling distribution.

Ok, but we don’t know μ, σ or the shape of the population distribution, so we don’t know exactly what the sampling distribution is.

If we did, we wouldn’t be asking a question about the population.

A common way to proceed is to assume the sampling distribution is Normal.
The Normal distribution

A particular distribution shape.
Defined by a mathematical function.
Completely specified by it’s mean (center) and standard deviation (spread).

Useful approximation to many distributions, but, very few things are exactly Normal.

**68-95-99.7% rule:**
If data is Normally distributed, 68% of observations will be within 1 standard deviation of the mean, 95% within 2 sds, 99.7% within 3 sds.
Your turn

Which of these are Normal distributions?
Next time

Use the facts about the sampling distribution for the **sample average**, to construct a range of likely values for the **population mean**.

Did today’s material feel foreign?

Read Chapter 4 in OpenIntro: